1258413

THE UNITED STATES OF AMERICA

TO ALL TO WHOM THESE PRESENTS SHALL COME:

UNITED STATES DEPARTMENT OF COMMERCE

United States Patent and Trademark Office

December 08, 2004

THIS IS TO CERTIFY THAT ANNEXED HERETO IS A TRUE COPY FROM THE RECORDS OF THE UNITED STATES PATENT AND TRADEMARK OFFICE OF THOSE PAPERS OF THE BELOW IDENTIFIED PATENT APPLICATION THAT MET THE REQUIREMENTS TO BE GRANTED A FILING DATE.

APPLICATION NUMBER: ***60/518,220***
FILING DATE: ***November 07, 2003***
RELATED PCT APPLICATION NUMBER: ***PCT/US04/37291***

Certified by

Jon W Dudas

Acting Under Secretary of Commerce
for Intellectual Property
and Acting Director of the U.S.
Patent and Trademark Office

UNITED STATES PATENT AND TRADEMARK OFFICE
DEPARTMENT OF COMMERCE

PATENT

Express Mail Label No.: EL 992 785 426 US

Attorney Docket No.: DNACOMP-08434

PROVISIONAL APPLICATION FOR PATENT COVER SHEET

This is a request for filing a PROVISIONAL APPLICATION FOR PATENT under 37 C.F.R. 1.53(b)(2).

Docket Number	DNACOMP-08434	Type a plus sign (+) inside this box →	

INVENTOR(s) / APPLICANT(s)

Last Name	First Name	Middle Initial	Residence (City and Either State or Foreign Country)
SantaLucia	**John**		**Grosse Pointe Woods, MI**
Hicks	**Donald**	**A.**	**Ann Arbor, MI**

TITLE OF THE INVENTION (280 Characters Max.)

System and Methods for Three Dimensional Molecular Structural Analysis

CORRESPONDENCE ADDRESS

MEDLEN & CARROLL, LLP
101 Howard Street, Suite 350
San Francisco, California 94105
United States of America

ENCLOSED APPLICATION PARTS *(Check All That Apply)*

X	Specification	Number of Pages	47	___	Small Entity Statement
X	Drawing(s)	*Number of Sheets*	0	___	Other (Specify): Power of Attorney
				___	Other (Specify): Assignment

METHOD OF PAYMENT OF FILING FEES FOR THIS PROVISIONAL APPLICATION FOR PATENT

___	Charge Account No. **08-1290** in the amount of **$80.00. An originally executed duplicate of this transmittal is enclosed for this purpose.**	FILING FEE AMOUNT ($)	**$80.00**
X	The Commissioner is hereby authorized to charge any deficiency in the payment of the required fee(s), and/or credit any overpayment, to Deposit Account No.: **08-1290. An originally executed duplicate of this transmittal is enclosed for this purpose.**		

This invention was made by an agency of the United States Government under a contract with an agency of the United States Government.

X No.
___ Yes, the name of the U.S. Government agency and the Government contract number are: ______________

Respectfully submitted,

Date: November 7, 2003

Mary Ann D. Brow
Reg. No.: 42,363

MEDLEN & CARROLL, LLP
101 Howard Street, Suite 350
San Francisco, California 94105
608/218-6900

___ Additional inventors are being named on separately numbered sheets attached hereto.

PATENT

Express Mail Label No.: **EL 992 785 426 US** Attorney Docket No.: **DNACOMP-08434**

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re application of: John SantaLucia, *et al.*

For: **System and Methods for Three Dimensional Molecular Structural Analysis**

Mail Stop Provisional Patent Application
Commissioner for Patents
P.O. Box 1450
Alexandria, Virginia 22313-1450

CERTIFICATION UNDER 37 C.F.R. § 1.10

I hereby certify that this correspondence and the documents referred to as attached therein are being deposited with the United States Postal Service on **November 7, 2003**, in an envelope as "EXPRESS MAIL POST OFFICE TO ADDRESSEE" service under 37 C.F.R. § 1.10, Mailing Label Number **EL 992 785 426 US**, addressed to: **Mail Stop Provisional Patent Application**, Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.

Susan M. McClintock
Susan M. McClintock

TRANSMITTAL COVER SHEET FOR FILING PROVISIONAL APPLICATION (37 C.F.R. § 1.51(2)(i))

This is a request for filing a PROVISIONAL APPLICATION FOR PATENT under 37 C.F.R. 1.53(b)(2).

1. The following comprises the information required by 37 C.F.R. § 1.51(a)(2)(i)(A):

2. The name(s) of the inventor(s) is/are (37 C.F.R. § 1.51(a)(2)(i)(B)):

   John SantaLucia
   Donald A. Hicks

3. Address(es) of the inventor(s), as numbered above (37 C.F.R. § 1.51(a)(2)(i)(C)):

   563 North Rosedale Court, Grosse Pointe Woods, MI 48236
   233 Hunters Trail, Ann Arbor, MI 48103

4. The title of the invention is (37 C.F.R. § 1.51(a)(2)(i)(D)):

   System and Methods for Three Dimensional Molecular Structural Analysis

5. The name, registration, and telephone number of the attorney (*if applicable*) is (37 C.F.R. § 1.51(a)(2)(i)(E)):

Mary Ann D. Brow
Reg. No.: 42,363
Tel.: 608/218-6900

(complete the following, if applicable)

___ A Power of Attorney accompanies this cover sheet.

6. The docket number used to identify this application is (37 C.F.R. § 1.51(a)(2)(i)(F)):

Docket No.: **DNACOMP-08434**

7. The correspondence address for this application is (37 C.F.R. § 1.51(a)(2)(i)(G)):

MEDLEN & CARROLL, LLP
101 Howard Street, Suite 350
San Francisco, California 94105

8. Statement as to whether invention was made by an agency of the U.S. Government or under contract with an agency of the U.S. Government. (37 C.F.R. § 1.51(a)(2)(i)(H)):

This invention was made by an agency of the United States Government, or under contract with an agency of the United States Government.

X No.

___ Yes.

The name of the U.S. Government agency and the Government contract number are: ________.

9. Identification of documents accompanying this cover sheet:

A. Documents required by 37 C.F.R. § 1.51(a)(2)(ii)-(iii):

Specification:	No. of pages	47
Drawings:	No. of sheets	0

B. Additional documents:

X Claims: No. of claims 8

___ Power of Attorney

___ Small Entity Statement

___ Assignment

___ Other

10. Fee

The filing fee for this provisional application, as set in 37 C.F.R. § 1.16(k), is $160.00 for other than a small entity, and $80.00 for a small entity.

X Applicant is a small entity.

11. Small Entity Statement

___ The verified statement(s) that this is a filing by a small entity under 37 C.F.R. §§ 1.9 and 1.27 is(are) attached.

12. Fee payment being made at this time

___ Charge Account No. **08-1290** in the amount of **$80.00. An originally executed duplicate of this transmittal is enclosed for this purpose.**

13. Method of Fee Payment:

X Check in the amount of **$80.00**

___ Charge Account No. **08-1290** in the amount of **$80.00. A duplicate of this Cover Sheet is attached.**

X Please charge Account No. **08-1290** for any fee deficiency. **A duplicate of this Cover Sheet is attached.**

Date: November 7, 2003

Mary Ann D. Brow
Reg. No.: 42,363

MEDLEN & CARROLL, LLP
101 Howard Street, Suite 350
San Francisco, California 94105
608/218-6900

EL992785426US

SYSTEM AND METHODS FOR THREE DIMENSIONAL MOLECULAR STRUCTURAL ANALYSIS

FIELD OF THE INVENTION

The present invention relates to methods and systems for the accurate prediction of nucleic acid (*e.g.*, RNA and DNA) and other macromolecular three-dimensional structure from sequence and constraint information.

BACKGROUND OF THE INVENTION

The structures formed by macromolecules are generally essential to their function. For example, tRNA structure is critical to its proper function in being recognized by the cognate tRNA synthetase and binding to the ribosome and correct mRNA codon, ribosomal RNA (rRNA) structures are essential to the correct function of the ribosome, and correct folding is essential to the catalytic function of Group I self-splicing introns (*See e.g.*, the chapters by Woese and Pace (p. 91), Noller (p. 137), and Cech (p. 239) *in* Gesteland and Atkins (eds.), *The RNA World,* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY [1993]). Folded structures in viral RNAs have been linked to infectivity (Proutski *et al.*, J Gen Virol., 78(Pt 7):1543-1549 [1997], altered splicing (Ward, *et al.*, Virus Genes 10:91 [1995]), translational frameshifting (Bidou *et al.*, RNA 3:1153 [1997]), packaging (Miller, *et al.* J Virol., 71:7648 [1997]), and other functions. In both prokaryotes and eukaryotes, RNA structures are linked to post-transcriptional control of gene expression through mechanisms including attenuation of translation (Girelli *et al.*, Blood 90:2084 [1997], alternative splicing (Howe and Ares, Proc. Natl. Acad. Sci. USA 94:12467 [1997]) and signaling for RNA degradation (Veyrune *et al*, Oncogene 11:2127 [1995]). Messenger RNA secondary structure has also been associated with localization of that RNA within cells (Serano and Cohen, Develop., 121:3809-3818 [1995]). In DNA, it has been shown that cruciform structures have been tied to control of gene expression (Hanke *et al.*, J. Mol. Biol., 246:63 [1995]). It can be seen from these few examples that the use of folded structures as signals within

organisms is not uncommon, nor is it limited to non-protein-encoding RNAs, such as rRNAs, or to non-protein-encoding regions of genomes or messenger RNAs.

Rapid determination of nucleic acid structure would be a useful tool for basic and clinical research and for diagnostics. Accurate identification of nucleic acid structures would facilitate the design and application of therapeutic agents targeted directly at nucleic acids and related molecules.

Methods for the experimental determination of nucleic acid and other macromolecular structure (*e.g.*, NMR and X-ray crystallography), cannot keep pace with the exponential growth of databases of the primary sequences of such macromolecules. Thus, there is a need to develop tools for the prediction of structure from directly from sequence information.

SUMMARY OF THE INVENTION

The present invention relates to methods and systems for the accurate prediction of nucleic acid (*e.g.*, RNA and DNA, and other biomolecular mimics) three-dimensional structure from sequence and constraint information. In preferred embodiments, the test sequence that is analyzed by the systems and methods of the present invention is DNA or RNA. In some embodiments, the systems are automated and are used to generate large numbers of three-dimensional structures from sequences stored in databases (*e.g.*, public and private nucleic acid sequence databases). A wide variety of applications of the invention exist, including, but not limited to, basic research applications, diagnostic applications, therapeutic applications, and drug screening applications. In addition, the systems and methods of the present invention allow for rational design of folded nucleic acid molecules (with and without other associated molecules and ions), to generate novel materials, catalysts, and nanotechnologies.

For example, the present invention provides systems and methods for generating corrected three-dimensional models of nucleic acids and other biomolecules. By "corrected" it is intended to mean that initial models are generated (*e.g.*, using traditional modeling techniques or techniques of the present invention) and that the initial models are improved (made more accurate) by the systems and methods of the present invention.

In preferred embodiments, the systems comprise a processor that is configured to carry out one or more of the following tasks:

a) generate an initial, uncorrected model of a test sequence (*e.g.*, an sequence with unknown structure, a sequence with partially known structure, etc.) by comparison to a reference sequence (*i.e.*, a sequence with known structure);

b) align secondary structure constraints (see Detailed Description) of the reference sequence with the test sequence to generate an aligned sequence;

c) make substitutions, deletions, and insertions dictated by the aligned sequence (see Detailed Description) using geometrical computation algorithms for the substitutions and using molecular mechanics, molecular dynamics algorithms, and other algorithms (such as the Discrete Sampling of Torsion Angles with Rigid-body Rotations and Optimization) to close gaps caused by the deletions and insertions;

d) identify conserved hydrogen bonds present in both the reference sequence and the uncorrected model to select hydrogen bond constraints; and

e) optimize the uncorrected model using a forcefield algorithm (see Detailed Description) that accounts for (*i.e.*, takes into consideration) the hydrogen bond constraints to generate a corrected three-dimensional model of the test sequence.

The present invention also provides methods that employ the system to generate corrected three-dimensional models of test sequences.

The present invention further provides systems and methods for predicting biomolecular three-dimensional structure, either *de novo* or where some structural information is known about a test sequence that is to be analyzed. In some such embodiments, the present invention provides systems for predicting nucleic acid three-

dimensional structure, where the system comprises a processor configured to carry out one or more of the following tasks:

a) compute one or more secondary structures of a test nucleic acid (*e.g.*, using standard secondary structure prediction methods known in the art and/or those described herein);

b) decompose the secondary structures into nucleic acid structure motifs (*e.g.*, base pairs, hairpins, bulges, internal loops, etc.);

c) rank the structure motifs in a hierarchal tree (*i.e.*, an organizational structure that prioritizes motifs by category and defines an interrelationship between the different categories of motifs—*e.g.*, hairpins as the "leaves", internal loops and bulges as the "branches", and multiloops and bifurcations as sub-roots and roots, respectively);

d) identify candidate three-dimensional motif structures for the motifs from a database of known three-dimensional structure motifs (*e.g.*, for each motif, select similar structures from a motif structure database);

e) link the candidate three-dimensional motif structures in an order specified by the hierarchal tree to generate a candidate three-dimensional composite structure;

f) submit the candidate three-dimensional composite structure to an energy minimization algorithm (*e.g.*, AMBER, CHARMM, the methods of the present invention, etc.) to generate a refined candidate three-dimensional structure/s;

g) select a refined candidate three-dimensional structure based on best calculated energy (*e.g.*, selecting the structure, among many that are generated, with the lowest total energy) to predict a three-dimensional structure of the test nucleic acid.

In some embodiments, the candidate three-dimensional motif structures comprise known secondary structure elements (*e.g.*, known from phylogeny [*e.g.*, comparative

sequence analysis], known from experimental methods [*e.g.*, site directed mutagenesis, chemical probing, nuclease probing, etc.], etc.). In some preferred embodiments, the refined candidate three-dimensional structure is selected by ranking a plurality of such structures by total energy defined by the sum of forcefield energy (see Detailed Description) and secondary structure folding energy from a dynamic programming algorithm (*e.g.*, Visual OMP—see below). The present invention also provides methods for generating a three-dimensional structure of test nucleic acids using such systems.

The present invention also provides systems and methods for generating and managing biomolecule structure motif databases (*e.g.*, for use in the methods above). For example, the present invention provides systems comprising a processor that is configured to carry out one or more of the following tasks:

a) receive nucleic acid physical structure information (*e.g.*, sequence, secondary structure, etc.);

b) decompose said physical structure information into nucleic acid structure motifs;

c) automatically determine hydrogen bonds, base pairs, mismatches, and a variety of folding motifs (stacking, U-turns, A-platforms, coaxial stacking, etc.)

d) associate data with said structure motifs, said data comprising one or more of: type of motif (hairpin, bulge, internal loop, multiloop, mismatch, coaxial stack), size of motif, coordinates of backbone (*e.g.*, xyz coordinates, ribose-phosphate for RNA, deoxyribose-phosphate for DNA, modified backbone for modified nucleic acids), source of the coordinates (*e.g.*, Protein Data Bank accession number), reliability parameter (*i.e.*, the resolution, a score or indication of the reliability of the source of the information), sequences known to form a motif, and dihedral angles for bases;

e) compare the nucleic acid structure motifs to existing motifs in the database; and

f) add the structure motif and associate data to said database.

In preferred embodiments, the coordinates in the database are derived from NMR and X-ray structures, other experimental techniques (*e.g.*, cryo-electron microscopy, atomic force microscopy, fluorescence resonance energy transfer), or previously predicted structures from the invention (*e.g.*, in some embodiments the invention can "learn" from each example for which it is used), or other sources, or are obtained from databases or literature references. In some embodiments, the comparison is carried out by determining the root mean squared deviation of the new motif to the existing motifs to determine whether the motif is unique in the database. In some embodiments, whether the new motif is unique to the database or not, the motif is still cataloged in the database. The present invention also provides methods for generating a nucleic acid structure motif database using the above system.

The present invention further provides systems and methods for refining nucleic acid structure predictions. For example, the present invention provides a system for refining nucleic acid structure predictions comprising a processor configured to carry out one or more of the following tasks:

a) calculate energy minimization terms for a test nucleic acid structure prediction model, said energy minimization terms comprising: bond stretching, bond angles, torsion stress, and non-bonded interactions (*e.g.*, van der Waals, hydrogen bond, electrostatic);

b) optimize force constants, distance dependence, partial charges, and van der Waals radii parameters;

c) account for gap penalties for insertions or deletions, if present in the prediction model (*e.g.*, present because of the use of methods of the present invention);

d) account for one or more experimental constraints associated with said test nucleic acid, said experimental constraints comprising hydrogen bonding information, nuclear Overhauser effect information, low resolution cryo-electron microscopy information, and chemical probing information;

e) employ distance constraints within a defined distance range but ignore distance constraints outside of said defined distance range; and

f) account for one or more nucleic acid folding thermodynamic measures, said nucleic acid folding thermodynamic measures comprising: folding entropy changes and solvation entropy changes as well as enthalpy and free energy changes at different temperature, salt, and other solution conditions.

g) Account for known interactions with proteins and metal ions by setting "anchor points" (see Detailed Description).

In some embodiments, the folding entropy and solvation entropy are obtained from solution measurements (*e.g.*, UV absorbance melting curves and calorimetric measurements known in the scientific literature) and decomposed into motif parameters (*e.g.*, parameters for base pairs, mismatches, and loops of various sizes such as hairpins, internal loops, multiloops, and bulges). The present invention also provides methods for refining a nucleic acid structure prediction using such systems.

The present invention also provides systems and methods for the identification of conserved hydrogen bonds for use in generating multi-dimensional biomolecule structure prediction. The systems and methods employ one or more of the following steps:

a) identification of all hydrogen bonds in a nucleic acid structure (*e.g.*, using prediction algorithms know in the art or described herein);

b) analysis of identified hydrogen bonds to determine if the nucleotides participate in Watson-Crick base pairs, mismatches, base-backbone interactions, and/or backbone/backbone interactions;

c) analyze the co-planarity of bases to confirm the presence of Watson-Crick base pairs; and

d) compare hydrogen bonds between reference sequences and test sequences to identify correct hydrogen bonds in the test sequences.

DEFINITIONS

As used herein, the term "nucleic acid" refers to strands comprising backbones (*e.g.*, of ribose phosphate and deoxyribose phosphate) and side chains generally comprising heterocyclic bases such as A, C, G, T, and U. Nucleic acids comprise "natural" nucleic acids, *i.e.*, those comprising natural backbones of ribose phosphate and deoxyribose phosphate, and side chains comprising the most common heterocyclic bases: A, C, G, T, and U. Examples of natural nucleic acids include deoxyribonucleic acid (DNA) and ribonucleic acid (RNA).

As the term is used herein, nucleic acids also comprise synthetic analogs of DNA or RNA in which the backbone and/or base moieties are substituted. Examples of synthetic nucleic acids include but are not limited to PNA (Nielsen PE, *et al.*, Science 254 (5037): 1497-1500 Dec. 6 1991), LNA (Petersen M., *et al.*, Nucleosides Nucleotides & Nucleic Acids 22 (5-8): 1691-1693 2003), TNA (Chaput JC, Szostak JW (2003) J Am Chem Soc 125 (31): 9274-9275), 2'-O-methyl-RNA (Schubert S, *et al.*, (2003) Nucleic Acids Res. 31 (20): 5982-5992), MOE (Vickers TA, *et al.*, Nucleic Acids Res. 29 (6): 1293-1299 Mar. 15 2001), 2'-fluoro (Shimizu M., *et al.*, FEBS Letters 302 (2): 155-158 May 11 1992), Hexose (Eschenmoser A, Hexose Nucleic-Acids, Pure Appl Chem 65 (6): 1179-1188 June 1993), 3-nitro-pyrrole, 5-nitro-indol, etc.

Nucleic acids may be single stranded, double stranded or may comprise both single and double stranded regions. Nucleic acids may comprise unimolecular folds, may comprise duplexes with strands of equal length or, in the case of a short oligo hybridizing to a long oligonucleotide, for example, may comprise an intermolecular duplex and tails that fold to form an intramolecular duplex or other structure. Nucleic acids may also comprise multiple stranded structures in which more than two strands of nucleic acid associate to form a higher order structure.

The terms "analog" and "modified" are used interchangeably herein in reference to bases, nucleosides and nucleotides other than the most common bases and nucleotides, *i.e.*, A, T, G, C, and U. Such analogs and modified bases and nucleotides include modified natural nucleotides and non-naturally occurring nucleotides, including but not limited to N4-acetyl-C, 5-methyl-C, m4Cm, inosine, 2-thio-U, dihydro-U, pseudo-U, N3-methyl-pseudo-U, N3-methyl-U, 4-thio-U, rT, Y-base, 2'-O-methyl A, C, G, U, N2-

methyl-G, N7-methyl-G, N6-methyl-A, N6-dimethyl-A. At least 95 naturally occurring modifications have been observed in RNA and DNA (Rozenski J, *et al.*, The RNA modification database: 1999 update, Nucleic Acids Res 27 (1): 196-197 JAN 1 1999). Analogs and modified nucleotides include those that have altered stacking interactions, such as 7-deaza purines (*i.e.*, 7-deaza-dATP and 7-deaza-dGTP); base analogs with alternative hydrogen bonding configurations (*e.g.*, such as iso-C and iso-G and other non-standard base pairs described in U.S. Patent No. 6,001,983 to S. Benner, and the selectively binding base analogs described in U.S. Patent No. 5,912,340 to Igor V. Kutyavin, *et al.*); non-hydrogen bonding analogs (*e.g.*, non-polar, aromatic nucleoside analogs such as 2,4-difluorotoluene, described by B.A. Schweitzer and E.T. Kool, J. Org. Chem., 1994, 59, 7238-7242, B.A. Schweitzer and E.T. Kool, J. Am. Chem. Soc., 1995, 117, 1863-1872); "universal" bases such as 5-nitroindole and 3-nitropyrrole; and universal purines and pyrimidines (such as "K" and "P" nucleotides, respectively; P. Kong, *et al.*, Nucleic Acids Res., 1989, 17, 10373-10383, P. Kong *et al.*, Nucleic Acids Res., 1992, 20, 5149-5152). Nucleotide analogs include modified forms of deoxyribonucleotides as well as ribonucleotides, as well as those comprising sugars other than ribose.

As used herein, the term "pairing" in reference to nucleotides or nucleic acid strands refers to interaction between nucleotides or nucleic acid strands by the formation of hydrogen bonds. Pairing comprises thermodynamically favorable "Watson-Crick" pairs (*i.e.*, G-C and A-T pairs in DNA and G-C and A-U pairs in RNA. Pairing also comprises non Watson Crick "mismatch" pairs. G-T pairs in DNA and G-U pairs in RNA, referred to as "wobble pairs", are stable mismatches. Other mismatches include but are not limited to: G-G, G-A, A-A, T-T, C-C, C-T, A-C (in approximate decreasing order in stability in DNA, Peyret ref). A similar order applies to RNA, with U replacing T. Pairing also comprises natural base: analog pairs and analog-analog pairs.

As used herein, the term "primary structure" refers to the sequential order of units in a strand or chain. As used in reference to nucleic acids, the primary structure is the sequence of nucleotides in the nucleic acid strand. As used in reference to a protein, the primary structure refers to the sequence of amino acids in the chain.

As used herein, the term "secondary structure" refers to the representation of the pairing interactions between nucleotides, including pairing in pseudoknots. Secondary structure may be represented in a number of ways. For example, it may be represented in two dimensions, *e.g.*, on a Nussinov circle plot, in which a nucleotide sequence is mapped on a circle and pairing interactions are denoted by chords, or it may be drawn with the paired nucleotides close to one another (*e.g.*, as shown after Step 2 in Figure 2, for a tRNA). Secondary structure may also be represented in three dimensions, *e.g.*, as in a fluctional 3D structure of a nucleic acid in which the pairs adopt the A-form or B-form helical structures, but the loop nucleotides are partially disordered (*e.g.*, populated according to a Boltzmann distribution), and the relative orientations of helices with respect to one another is not specified exactly. The 3D representation is used when referring to the hierarcal folding process of nucleic acids.

As used herein, the term "pseudoknot" refers to any structure wherein, when mapped on Nussinov circle plot, there is crossing of the chords that denote pairing interactions.

Reference is made to different lengths of nucleic acids, *e.g.*, they may be characterized as large or long, medium, and small or short. As used herein, "small" or "short" means less than 25 nucleotides in length; "medium" means 25 to 100 nucleotides in length, and "large" or "long" means greater than 100 nucleotides in length.

As used herein, the term "constraint" refers to an aspect of a structure that might otherwise be variable, but that is assigned a particular value (*e.g.*, a property, position or relationship) during modeling of a structure. Constraints may comprise experimental or theoretically derived aspects of a structure, including but not limited to: distances between components of a structure, (*e.g.*, from NMR NOE measurements or FRET measurements); dihedral angles (*e.g.*, from NMR J-coupling measurements); directions with respect to an axis (*e.g.*, from NMR residual dipolar coupling measurements); exposure of a component to the surface or a structure (as determined by, *e.g.*, EDTA-Fe probing), exposure to solvent (as determined by, *e.g.*, reaction with DMS, DEPC, ENU, CMCT or kethoxal reagents); positions of nucleotides (as determined by, *e.g.*, low resolution X-ray crystallography, cryo-electron microscopy, atomic force microscopy, or NMR methods); other aspects of nucleotide disposition in a structure (*e.g.*, proximity to

other nucleotides, paired or unpaired status, or pairing with a particular other nucleotide) such as can be determined by, for example, cross-linking [*e.g.*, using psoralin or mustard reagents) or nuclease sensitivity (*e.g.*, Nucleases S1 and V1, or structure-specific nucleases such as FENs).

As used herein, the terms "processor" and "central processing unit" or "CPU" are used interchangeably and refer to a device that is able to read a program from a computer memory (*e.g.*, ROM or other computer memory) and perform a set of steps according to the program.

As used herein, the terms "computer memory" and "computer memory device" refer to any storage media readable by a computer processor. Examples of computer memory include, but are not limited to, RAM, ROM, computer chips, digital video discs (DVD), compact discs (CDs), hard disk drives (HDD), and magnetic tape.

As used herein, the term "computer readable medium" refers to any device or system for storing and providing information (*e.g.*, data and instructions) to a computer processor. Examples of computer readable media include, but are not limited to, DVDs, CDs, hard disk drives, magnetic tape and servers for streaming media over networks.

As used herein, the term "encode" refers to the process of converting one type of information or signal into a different type of information or signal to, for example, facilitate the transmission and/or interpretability of the information or signal. For example, image files can be converted into (*i.e.*, encoded into) electrical or digital information. Likewise, light patterns can be converted into electrical or digital information that provides and encoded video capture of the light patterns.

As used herein, the term "hyperlink" refers to a navigational link from one document to another, or from one portion (or component) of a document to another. Typically, a hyperlink is displayed as a highlighted word or phrase that can be selected by clicking on it using a mouse to jump to the associated document or documented portion.

As used herein, the term "Internet" refers to any collection of networks using standard protocols. For example, the term includes a collection of interconnected (public and/or private) networks that are linked together by a set of standard protocols (such as TCP/IP, HTTP, and FTP) to form a global, distributed network. While this term is intended to refer to what is now commonly known as the Internet, it is also intended to

encompass variations that may be made in the future, including changes and additions to existing standard protocols or integration with other media (*e.g.*, television, radio, etc). The term is also intended to encompass non-public networks such as private (*e.g.*, corporate) Intranets.

As used herein, the terms "World Wide Web" or "web" refer generally to both (i) a distributed collection of interlinked, user-viewable hypertext documents (commonly referred to as Web documents or Web pages) that are accessible via the Internet, and (ii) the client and server software components which provide user access to such documents using standardized Internet protocols. Currently, the primary standard protocol for allowing applications to locate and acquire Web documents is HTTP, and the Web pages are encoded using HTML. However, the terms "Web" and "World Wide Web" are intended to encompass future markup languages and transport protocols that may be used in place of (or in addition to) HTML and HTTP.

As used herein, the term "web site" refers to a computer system that serves informational content over a network using the standard protocols of the World Wide Web. Typically, a Web site corresponds to a particular Internet domain name and includes the content associated with a particular organization. As used herein, the term is generally intended to encompass both (i) the hardware/software server components that serve the informational content over the network, and (ii) the "back end" hardware/software components, including any non-standard or specialized components, that interact with the server components to perform services for Web site users.

As used herein, the term "HTML" refers to HyperText Markup Language that is a standard coding convention and set of codes for attaching presentation and linking attributes to informational content within documents. During a document authoring stage, the HTML codes (referred to as "tags") are embedded within the informational content of the document. When the Web document (or HTML document) is subsequently transferred from a Web server to a browser, the codes are interpreted by the browser and used to parse and display the document. Additionally, in specifying how the Web browser is to display the document, HTML tags can be used to create links to other Web documents (commonly referred to as "hyperlinks").

As used herein, the term "HTTP" refers to HyperText Transport Protocol that is the standard World Wide Web client-server protocol used for the exchange of information (such as HTML documents, and client requests for such documents) between a browser and a Web server. HTTP includes a number of different types of messages that can be sent from the client to the server to request different types of server actions. For example, a "GET" message, which has the format GET, causes the server to return the document or file located at the specified URL.

As used herein, the term "URL" refers to Uniform Resource Locator that is a unique address that fully specifies the location of a file or other resource on the Internet. The general format of a URL is protocol://machine address:port/path/filename. The port specification is optional, and if none is entered by the user, the browser defaults to the standard port for whatever service is specified as the protocol. For example, if HTTP is specified as the protocol, the browser will use the HTTP default port of 80.

As used herein, the term "PUSH technology" refers to an information dissemination technology used to send data to users over a network. In contrast to the World Wide Web (a "pull" technology), in which the client browser must request a Web page before it is sent, PUSH protocols send the informational content to the user computer automatically, typically based on information pre-specified by the user.

As used herein, the term "in electronic communication" refers to electrical devices (*e.g.*, computers, processors, NMR devices, fluorescent readers, etc.) that are configured to communicate with one another through direct or indirect signaling. For example, a conference bridge that is connected to a processor through a cable or wire, such that information can pass between the conference bridge and the processor, are in electronic communication with one another. Likewise, a computer configured to transmit (*e.g.*, through cables, wires, infrared signals, telephone lines, etc) information to another computer or device, is in electronic communication with the other computer or device.

As used herein, the term "transmitting" refers to the movement of information (*e.g.*, data) from one location to another (*e.g.*, from one device to another) using any suitable means.

GENERAL DESCRIPTION OF THE INVENTION

Two approaches have commonly been applied to elucidate nucleic acid secondary structures: physical approaches, such as analysis of crystal structure or NMR, and analytical approaches, such as comparative or phylogenetic analysis. Physical analysis remains the only way to get a complete determination of a folded structure for any given nucleic acid. However, that level of analysis is impractical if the goal is to analyze a large number of molecules. By far, the most often used method of analyzing biological nucleic acids is a phylogenetic, or comparative approach. This method of analysis is based on the biological paradigm that functionally homologous sequences will adopt similar structures. Sequences are screened for sequence conservation, stem-loop conservation, and for compensatory sequence changes that preserve predicted structures. Unfortunately, such analysis can only be applied when the number of related sequences is large enough for statistical analysis.

Methods for macromolecular structure determination (NMR, X-ray crystallography, cryo-electron microscopy, and atomic force microscopy) are labor and time intensive, and thus cannot keep pace with the exponential growth of naturally occurring sequence databases (*e.g.* GENBANK) or with synthetic sequences such as aptamers or rationally (by humans or by computers) designed sequences.

Thus, there is a need to develop tools for structure prediction from sequence. The systems and methods of the present invention provide means for accurately predicting nucleic acid (including, but not limited to, DNA, RNA, natural modifications, synthetic modifications and nucleic acid analogs) from sequence and constraint information. The systems and methods of the present invention find application in rational drug design (*e.g.*, design or selection of antibiotics against pathogen ribosomes, anticancer therapeutics targeted to telomerase, spliceosomes, and other RNA processing enzymes that are more active in cancerous cells than normal cells). In addition, the systems and methods of the present invention find application to the *in silico* design of nucleic acid-based structured nano-materials {Seeman, 1998 #294}, nano-robots or nano-machines (Yurke ref, Ehud Shapiro ref), and computing {Adelman, 1994 #296}.

Progress has been made on the "protein folding problem" in the past 30 years {, 2002 #144}. Nonetheless, it is still very challenging to accurately predict protein structure from sequence information alone and rational protein design is still in its infancy. In contrast, relatively little progress has been made on the RNA and DNA folding problems. The current state of the art for RNA and DNA secondary structure prediction is approximately 73%, and 85%, respectively {34, Mathews, 1999}{SantaLucia, 2003 #556}, and very little software is available for predicting DNA and RNA tertiary structure{35, Gautheret, 1993}.

Several groups have anticipated that prediction of RNA secondary structure may be considerably easier than the protein folding problem {Tinoco, 1999 #557;Turner, 1988 #140} (Draper ref), but general software for highly accurate automated 3D structure prediction of large, medium and small nucleic acids has not been reported in the literature. RNA has only 4 different residues all of which contain a heterocyclic aromatic base, while proteins have 20 different amino acids with diverse chemical functionality (apolar, charged, sulfhydryl, aromatic, etc.). In addition, RNA has strong pairing rules (G-C and A-U), while there are no such rules for proteins. These strong pairing rules result in well-defined secondary structure and domain boundaries that are readily predicted by comparative sequence analysis even when the sequence homology is low (which is not the case for proteins). Unfortunately, for an RNA of length N, there are approximately 1.8^N possible structures (M. Zuker & D. Sankoff. RNA Secondary Structures and their Prediction. *Bull. Mathematical Biology* 46, 591-621 (1984)); this makes the search for the global optimum impossible to determine for $N > 50$. Fortunately, the discrete nature of base pairing also makes RNA folding accessible to dynamic programming algorithms, which are very efficient – the global minimum is guaranteed to be found, along with structures ~~near the~~ optimum, with calculation time proportional to N^3, which is tractable for $N < 10{,}000$ with routinely available computers {36, Zuker, 1989}. The main drawback of using dynamic programming algorithms, however, is that RNA is represented as letters rather than three-dimensional atomic structures, which means that the folding rules are only approximate and incomplete, and they neglect "pseudoknot structures"{Pleij, 1995 #57}. In contrast, classical molecular dynamics simulations (such as AMBER and CHARMM) provide an essentially complete

atomic description and thus they are able to correctly converge on the correct structure, but only if started with a structure sufficiently close to the global optimum. The classical molecular dynamics simulations are not capable of widely searching conformation space, however.

The systems and methods of the present invention combine the strengths of dynamic programming algorithms and classical molecular dynamics simulations with novel algorithms for homology modeling, sequence alignment, nucleic acid geometrical manipulations, novel hybrid forcefield, and novel structural motif databases. In combination, the systems and methods of the present invention allow accurate prediction of nucleic acid structure from primary sequence information and constraint information.

A general overview of the forward and backward prediction methods of the present invention is summarized in Figure 1.

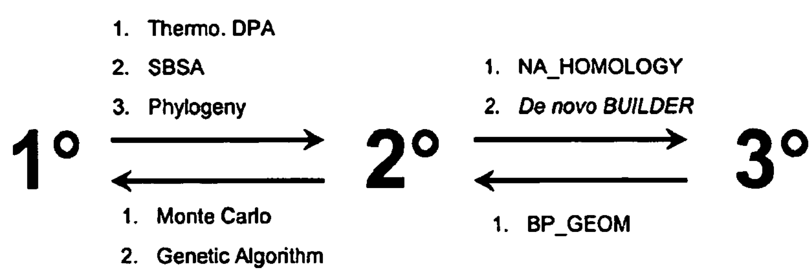


Figure 1: Overview of forward and reverse folding of nucleic acids, through primary (1°), secondary (2°), and tertiary (3°) or 3D structures.

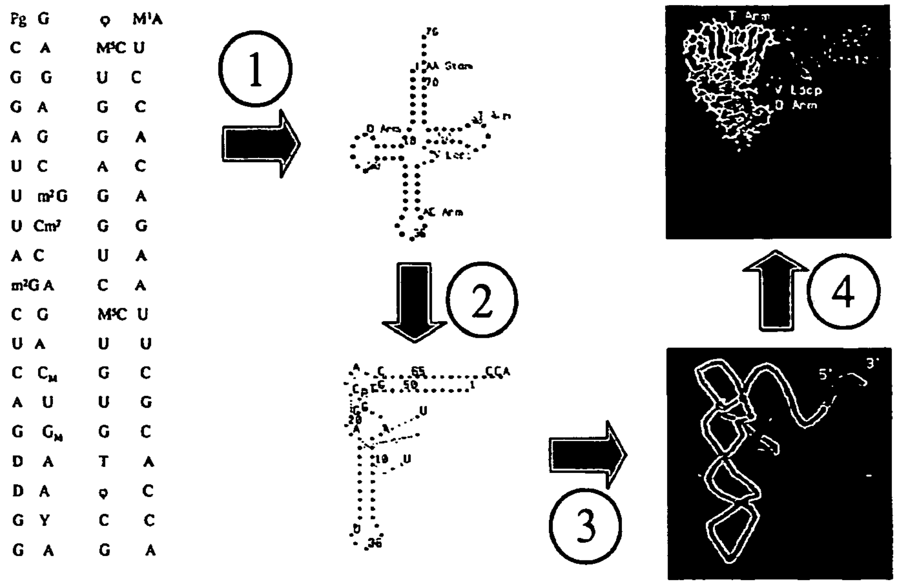


Figure 2. Overview of an approach to RNA structure prediction as diagrammed using tRNA molecule in some embodiments of the present invention. Step 1 is a dynamic programming algorithm for predicting the secondary structure from the sequence. Step 2 is an algorithm for predicting coaxial stacking of helices and pseudoknots. Step 3 is embeds the secondary structure in a 3D backbone fold. Step 4 creates a full-atom representation of the 3D (tertiary) structure and performs structure optimization by classical molecular dynamics and energy minimization algorithms.

DETAILED DESCRIPTION OF THE INVENTION

The systems and methods of the present invention combine the strengths of dynamic programming algorithms and classical molecular dynamics simulations with new methods for homology modeling, sequence alignment, nucleic acid geometrical manipulations, hybrid forcefield, and novel structural motif databases. These combinations provide new tools for the accurate prediction of nucleic acid structure from sequence and constraint information.

This Detailed Description of the Invention comprises the following sections: I: Homology modeling of nucleic acids; II: De novo 3D structure prediction of nucleic acids; III. Nucleic acid motif database; IV: An advanced forcefield for nucleic acids; V: V. Structure Based Sequence Alignment and Threading of Nucleic Acids; VI: Identification of Conserved Hydrogen bonds; VII: Systems of the Present Invention.

These descriptions are illustrations of certain preferred embodiments of the present invention are not intended to limit the scope thereof.

I. Homology modeling of nucleic acids.

For many functional RNAs (tRNA, rRNAs, ribozymes, etc.), the secondary structure may be accurately deduced by comparative sequence analysis {Gutell, 1992 #864} where at least one X-ray crystal or NMR structure representative of the class is available (referred to as the "reference template"). The NA_HOMOLOGY algorithm of the present invention "threads" a sequence whose three-dimensional structure is entirely or partially unknown (referred to as the "query sequence" of "test sequence") into the reference template coordinates. The first step performs a novel sequence alignment between the template and query sequences subject to secondary structure constraints present in the template structure. This is called "structure based alignment" (described in more detail below in Section V). The second step of the threading algorithm is to make the substitutions, deletions and insertions dictated by the sequence alignment using geometrical computations for the substitutions and using modified classical molecular mechanics and molecular dynamics to close the gaps caused by deletions and insertions (described in more detail below in Section III). Insertions may also be accomplished using the BUILDER algorithm (described in more detail below in Section II). The third step is to identify conserved hydrogen bonds (H-bonds) present in both the reference template and the initial homology model of the unknown sequence (described in more detail below in Section VI). The fourth step is to optimize the structure using classical forcefield methods (described in more detail below in Section IV) subject to the conserved H-bond constraints. This methodology has tremendous potential to, among other uses, leverage the genome sequencing projects by allowing the automated conversion of a large percentage of the functional RNAs in sequence databases into 3D structural model databases.

A summary of certain preferred embodiments of the systems and methods are shown below, in Figure 3.

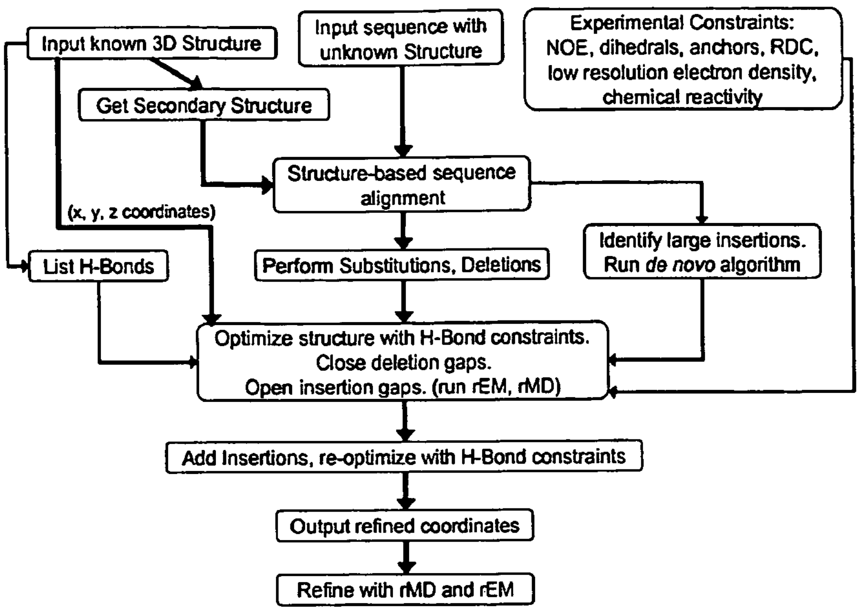


Figure 3: Flowchart of the NA_HOMOLOGY algorithm.

II. Method and Systems for three-dimensional (*e.g.*, de novo) structure prediction of nucleic acids.

Application of dynamic programming algorithms (DPA) for prediction of optimal and suboptimal secondary structures of RNA is well established in the literature {36, Zuker, 1989}. A thermodynamic DPA called Visual OMP (Oligonucleotide Modeling Platform) based on patent pending application number WO0194611 A2 WO (herein incorporated by reference in its entirety) can be used in the systems and methods of the present invention to compute optimal and suboptimal secondary structures of DNA and RNA and other modified nucleic acids. The BUILDER algorithm of the present invention converts each of the optimal and suboptimal secondary structures into several three-dimensional structures using an embedding algorithm. The BUILDER algorithm works in four steps: 1. the predicted secondary structure is decomposed into its constituent motifs (*e.g.*, base pairs, hairpins, bulges, internal loops, etc.) and used to generate a hierarchal tree (*e.g.*, with hairpins as the leaves, internal loops, bulges as the branches, and multiloops and bifurcations as sub-roots and roots, respectively), 2. candidate 3D

structures for each motif are retrieved from a novel motif database (see Section III), 3. the motifs are geometrically linked together in the order specified by the tree, and 4. classical molecular dynamics simulations and energy minimization using the forcefield (NA_FF; Section IV below) as well as AMBER or CHARMM or similar techniques used to refine the structures. If the secondary structure is known already from phylogeny (*e.g.*, comparative sequence analysis) or by experimental methods (*e.g.*, site directed mutagenesis, chemical probing, nuclease probing etc.), then the BUILDER algorithm can start with the correct secondary structure. The candidate structures are then be re-ranked according to a total energy that is a weighted sum of the forcefield energy and secondary structure folding energy from the dynamic programming algorithm (Visual OMP). The result is highly accurate three-dimensional *de novo* structure predictions of RNA, DNA, and modified nucleic acids or other biomolecules. Importantly, this approach is completely general and applies even to RNA and DNA sequences for which there are no available representative three-dimensional structures, which is tremendously useful for elucidating the structures of new functional RNAs discovered in genome projects or *in vitro* aptamer screens. This approach is also useful for *in silico* design of new RNA and DNA folds that do not exist in nature, but that have novel materials, catalytic, or nanotechnology applications.

A summary of certain preferred embodiments of the systems and methods are shown below, in Figures 4 and 5.

Figure 4: De novo Structure prediction algorithm.

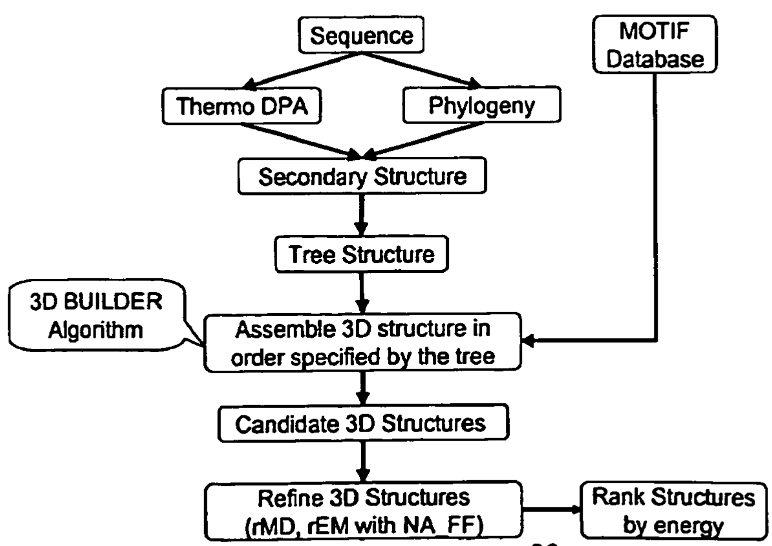
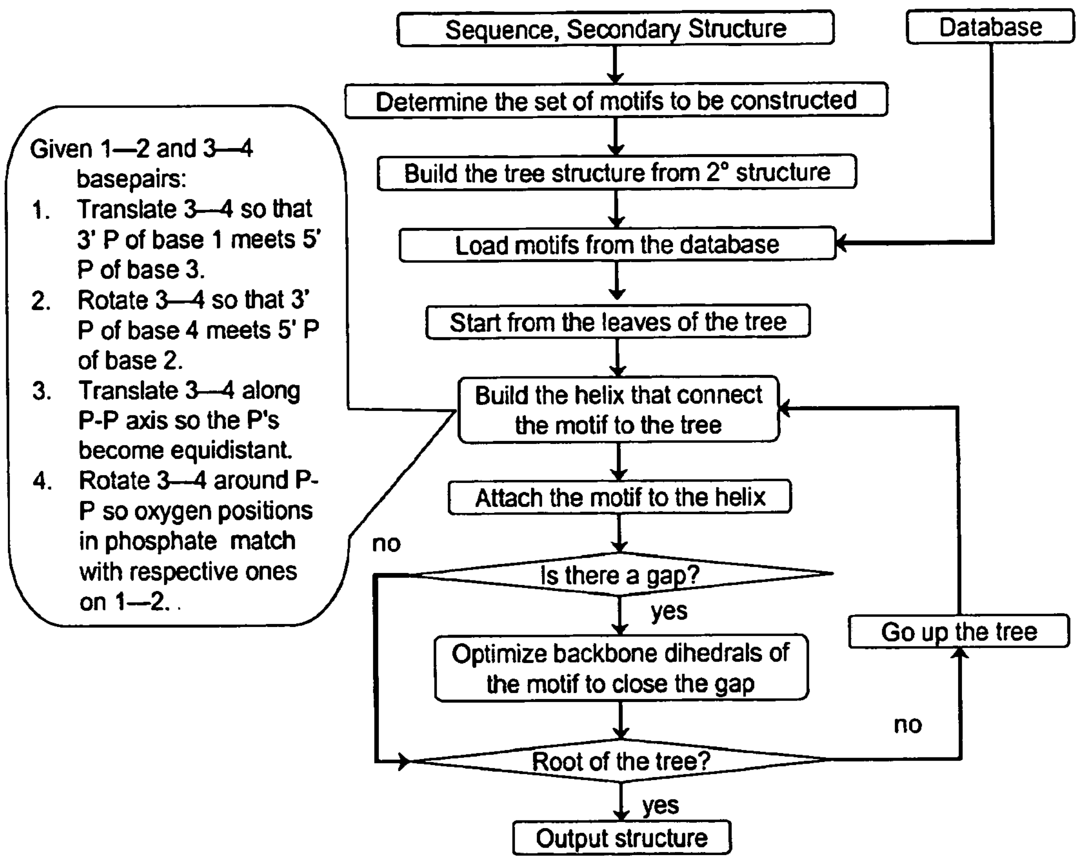


Figure 5: Flowchart of the BUILDER algorithm.



III. Nucleic acid motif database.

The flowchart describing the construction of a motif database is given in Figure 5. The database has been embodied in a flat file format and can be readily adapted to a relational database (an example entry is shown in figure 5A). In preferred embodiments, each entry in the database contains the keywords that describe the type of motif, size of the motif, source of the coordinates (*e.g.* PDB accession number), reliability parameter, sequences known to form the motif, and xyz coordinates of the backbone (ribose-phosphate for RNA, deoxyribose-phosphate for DNA, or modified backbone for modified nucleic acids), and dihedral angles for the base moiety of each of the nucleotides. Each motif contains the closing base pair or base pairs (for internal loops, bulges, and multiloops), so that it may be readily appended to the next base pair in a stem. The

coordinates in the database are derived from NMR and X-ray structures of larger DNAs and RNAs and modified nucleic acids that are found in the Protein Database (PDB). As new structures are added to the PDB, an algorithm called MOTIF automatically decomposes the structure into its motifs and compares the new motifs to those existing in the database to determine the root mean squared deviation (RMSD) with the existing motifs to determine if the new motif is unique (Figure 6). If the motif is not unique, the sequence of the motif may still be cataloged under the SEQUENCE keyword, so that it may be used for future sequence alignment searches (*e.g.*, for use in the methods described in Section V). Table 1 provides an example of the organization of data within a database of the present invention.

Table 1: Structured Motif Database

Hairpins[a,b,c,d,e]		Bulges[b,c,d,e]		Internal Loops[b,c,d,e]		Internal Loops[b,c,d,e]		Multiloop fragments[a,b,c,d,e]	
Length	Number	Length	Number	Length	Number	Length	Number	Length	Number
3	9	1	27	1x2	8	4x4	7	1	22
4	45	2	16	1x3	5	4x5	7	2	27
5	21	3	3	1x4	3	4x6	1	3	24
6	20	4	1	1x5	1	4x7	2	4	20
7	23	5	0	2x2	3	5x5	3	5	17
8	19	6	2	2x3	3	5x6	7	6	10
9	6	7	0	2x4	5	5x7	2	7	4
10	1			2x5	0	5x8	1	8	1
11	5			2x6	0	6x6	0	9	0
12	1			2x7	1	6x7	1	10	6
13	0			3x3	6	7x7	1	11	0
14	1			3x4	6	18x19	1	12	0
15	2			3x5	2			13	0
				3x6	3			14	3
				3x7	2			15	2
				3x9	1			16	1

References for crystal structures
a 1EHZ tRNAphe
b 1HR2 group I intron
c 1J5E 16S rRNA *Thermus thermophilus*
d 1JJ2 23S rRNA *Haloarcula marismortui*
e 1NKW 23S rRNA *Deinococcus Radiodurans*

Figure 5A. Example entry in the Motif database used by BUILDER.

```
Motif type:    Hairpin
Motif size:    3
Sequences:     CUCAG, CUAAG
Source:        1NKW
Positions      331-333
Resolution:    3.0
ATOM    6145  P      C 0 330       -42.207 137.234  85.023  1.00 48.85         P
ATOM    6146  O1P    C 0 330       -43.221 138.159  84.459  1.00 48.85         O
ATOM    6147  O2P    C 0 330       -41.551 136.236  84.137  1.00 48.85         O
ATOM    6148  O5*    C 0 330       -41.069 138.098  85.722  1.00 48.85         O
ATOM    6149  C5*    C 0 330       -41.357 138.842  86.901  1.00 48.85         C
ATOM    6150  C4*    C 0 330       -40.085 139.347  87.534  1.00 48.85         C
ATOM    6151  O4*    C 0 330       -39.275 138.237  88.009  1.00 48.85         O
```

```
ATOM   6152  C3*   C 0 330      -39.142 140.097  86.615  1.00 48.85          C
ATOM   6153  O3*   C 0 330      -39.580 141.430  86.387  1.00 48.85          O
ATOM   6154  C2*   C 0 330      -37.842 140.026  87.402  1.00 48.85          C
ATOM   6155  O2*   C 0 330      -37.815 140.934  88.486  1.00 48.85          O
ATOM   6156  C1*   C 0 330      -37.898 138.592  87.935  1.00 48.85          C
DIHEDRAL ANGLE = -3.10169

ATOM   6165  P     U 0 331      -39.228 142.152  84.994  1.00 49.32          P
ATOM   6166  O1P   U 0 331      -39.867 143.492  85.005  1.00 49.32          O
ATOM   6167  O2P   U 0 331      -39.524 141.207  83.886  1.00 49.32          O
ATOM   6168  O5*   U 0 331      -37.650 142.345  85.065  1.00 49.32          O
ATOM   6169  C5*   U 0 331      -37.025 142.775  86.271  1.00 49.32          C
ATOM   6170  C4*   U 0 331      -35.529 142.614  86.175  1.00 49.32          C
ATOM   6171  O4*   U 0 331      -35.164 141.210  86.081  1.00 49.32          O
ATOM   6172  C3*   U 0 331      -34.906 143.219  84.935  1.00 49.32          C
ATOM   6173  O3*   U 0 331      -34.775 144.623  85.078  1.00 49.32          O
ATOM   6174  C2*   U 0 331      -33.569 142.494  84.871  1.00 49.32          C
ATOM   6175  O2*   U 0 331      -32.611 143.025  85.763  1.00 49.32          O
ATOM   6176  C1*   U 0 331      -33.975 141.083  85.309  1.00 49.32          C
DIHEDRAL ANGLE = -2.90545

ATOM   6185  P     C 0 332      -35.030 145.572  83.810  1.00 49.17          P
ATOM   6186  O1P   C 0 332      -35.479 146.888  84.329  1.00 49.17          O
ATOM   6187  O2P   C 0 332      -35.875 144.848  82.823  1.00 49.17          O
ATOM   6188  O5*   C 0 332      -33.574 145.755  83.197  1.00 49.17          O
ATOM   6189  C5*   C 0 332      -32.912 144.694  82.509  1.00 49.17          C
ATOM   6190  C4*   C 0 332      -31.493 145.102  82.202  1.00 49.17          C
ATOM   6191  O4*   C 0 332      -31.535 146.504  81.860  1.00 49.17          O
ATOM   6192  C3*   C 0 332      -30.502 144.992  83.356  1.00 49.17          C
ATOM   6193  O3*   C 0 332      -29.831 143.732  83.217  1.00 49.17          O
ATOM   6194  C2*   C 0 332      -29.516 146.139  83.099  1.00 49.17          C
ATOM   6195  O2*   C 0 332      -28.392 145.752  82.336  1.00 49.17          O
ATOM   6196  C1*   C 0 332      -30.353 147.136  82.285  1.00 49.17          C
DIHEDRAL ANGLE = -1.83837

ATOM   6205  P     A 0 333      -29.305 142.921  84.508  1.00 48.22          P
ATOM   6206  O1P   A 0 333      -27.950 142.420  84.153  1.00 48.22          O
ATOM   6207  O2P   A 0 333      -30.352 141.955  84.920  1.00 48.22          O
ATOM   6208  O5*   A 0 333      -29.141 144.011  85.657  1.00 48.22          O
ATOM   6209  C5*   A 0 333      -27.920 144.727  85.823  1.00 48.22          C
ATOM   6210  C4*   A 0 333      -27.840 145.298  87.217  1.00 48.22          C
ATOM   6211  O4*   A 0 333      -29.030 146.099  87.451  1.00 48.22          O
ATOM   6212  C3*   A 0 333      -27.841 144.289  88.356  1.00 48.22          C
ATOM   6213  O3*   A 0 333      -26.521 143.812  88.625  1.00 48.22          O
ATOM   6214  C2*   A 0 333      -28.379 145.124  89.512  1.00 48.22          C
ATOM   6215  O2*   A 0 333      -27.397 145.973  90.068  1.00 48.22          O
ATOM   6216  C1*   A 0 333      -29.432 145.977  88.805  1.00 48.22          C
DIHEDRAL ANGLE = -2.16481

ATOM   6227  P     G 0 334      -26.302 142.545  89.599  1.00 47.04          P
ATOM   6228  O1P   G 0 334      -27.607 142.209  90.231  1.00 47.04          O
ATOM   6229  O2P   G 0 334      -25.131 142.841  90.464  1.00 47.04          O
ATOM   6230  O5*   G 0 334      -25.905 141.354  88.613  1.00 47.04          O
ATOM   6231  C5*   G 0 334      -26.903 140.481  88.083  1.00 47.04          C
ATOM   6232  C4*   G 0 334      -26.538 139.039  88.356  1.00 47.04          C
ATOM   6233  O4*   G 0 334      -27.732 138.225  88.372  1.00 47.04          O
ATOM   6234  C3*   G 0 334      -25.631 138.343  87.353  1.00 47.04          C
ATOM   6235  O3*   G 0 334      -24.275 138.645  87.643  1.00 47.04          O
ATOM   6236  C2*   G 0 334      -25.911 136.867  87.636  1.00 47.04          C
ATOM   6237  O2*   G 0 334      -25.129 136.345  88.692  1.00 47.04          O
ATOM   6238  C1*   G 0 334      -27.386 136.889  88.060  1.00 47.04          C
ATOM   6250  P     A 0 335      -23.277 139.110  86.472  1.00 46.60          P
ATOM   6251  O1P   A 0 335      -23.977 140.108  85.624  1.00 46.60          O
```

```
ATOM    6252   O2P    A 0 335      -22.691 137.896  85.844  1.00 46.60            O
ATOM    6253   O5*    A 0 335      -22.129 139.857  87.286  1.00 46.60            O
DIHEDRAL ANGLE = -0.820054
END
```

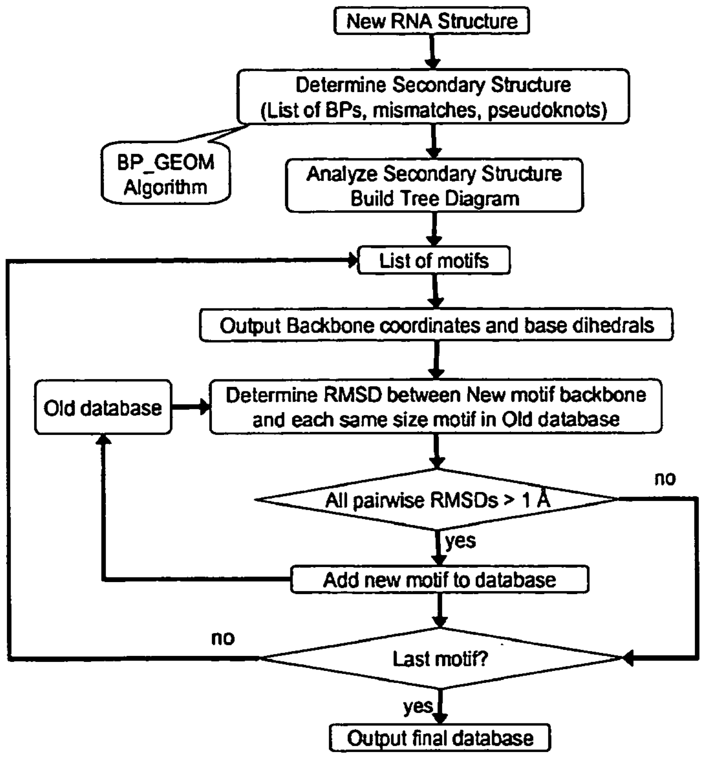


Figure 6: Algorithm for the MOTIF algorithm.

IV. An advanced forcefield for nucleic acids.

The present invention provides a novel forcefield specifically tailored to nucleic acid applications called "NA_FF". NA_FF includes the traditional terms for classical molecular dynamics and energy minimization including bond stretching, bond angles, torsion stress, and non-bonded interactions (van der Waals, H-bond, electrostatics). Specially optimized force constants, distance dependence, partial charges, and van der Waals radii parameters for nucleic acids are included as well as parameters for several modified nucleotides. NA_FF also includes pseudopotential terms for gap penalties (from insertions and deletions in Sections I and II, above), as well as pseudopotentials for

experimental constraints (H-bonds, NOEs, low resolution cryo-electron microscopy or X-ray electron densities, chemical modification data, etc.). A new optimization routine called "discrete sampling of torsion angles by rigid body rotations" is implemented by which whole segments of an RNA are rotated by discrete amounts (0.01 degree increments) for all dihedrals angles in a loop. For each discrete point (*i.e.*, torsion angle) the total energy function in evaluated (including the forcefield terms as well as pseudopotentials for gaps, and H-bond constraints). For each dihedral that best discrete torsion angle is kept. The procedure is repeated for all torsion angles until the energy function converges to a user-defined tolerance. These implementations allow for smooth minimization of ill-conditioned starting structures that contain large bond lengths (due to deletions) and overlapped atoms (due to insertions) that often cause traditional minimization algorithms to break chemical bonds, to not converge, or to crash (*e.g.* linmin failures). The constraint information is used in a novel "soft constraint" implementation in which distance constraints are imposed as parabolic or Lennard-Jones type function (with 6-12 or other exponents) in the desired distance range, but zero penalty outside the desired range; this implementation allows structures to widely search conformation space without getting stuck in very bad local minima and also accounts for potential inaccuracy in the constraint (*i.e.*, the constraint might be wrong and thus one would not want to apply a penalty for violating the constraint). An additional novel feature of the NA_FF forcefield is the ability to incorporate solution trends in DNA and RNA folding thermodynamics. Traditional forcefield methods do not account for folding entropy, solvation entropy, or residual loop entropy. These entropies are readily obtained from the solution measurements described in the literature and decomposed into motif parameters (*e.g.*, parameters for base pairs, mismatches and loops of various sizes such as hairpins, internal loops, multiloops, and bulges) (SantaLucia Turner refs). In addition, enthalpy values for base pairs and mismatches are exceptionally accurate and thus are also included in the NA_FF forcefield. These terms are added to the traditional forcefield terms (bond length, angle, non-bonded, etc) and weighted to avoid double counting interactions. This modified forcefield is used as part of a "total figure of merit energy" that is used to re-rank candidate three-dimensional structures (*e.g.*, for use in the systems and methods of Section II, above).

Ab initio quantum calculations have been performed on modified nucleotides to determine their geometries and RESP partial charges (Cornell ref 39). These results have been used to extend the NA_FF as well as the AMBER and CHARMM forcefields to include modified nucleotides. A novel "anchor points" approach is also implemented in which a given set of residues may be fixed in three-dimensional space by using a quadratic pseudopotential that penalizes the energy function if a structure during minimization moves an anchored residue away from its original position. Such anchoring allows for the orientation of intermolecular interactions of an RNA with other molecules to be retained. For example, tRNA is known to interact with its cognate synthetase protein through specific interactions and these interactions affect the shape of the tRNA – namely the angle of the "L" changes. Thus the effect of the synthetase can be model by freezing these interaction sites by declaring them to be anchor points. Such interactions between RNA and proteins can be generally modeled using anchors. The effects of metal that chelate RNA can also be modeled using anchors by freezing the positions of nucleotides known to interact with a magnesium or other metal observed in a crystal structure (or known from another experimental method). An additional example is the ribosomal RNAs, which interact with multiple proteins. The ribosomal proteins from different organisms are often quite divergent, but generally interact with RNA in similar fashion. Thus, the RNA component of a ribosome can be accurately modeled in different organisms, even if the protein structures in the reference and query organisms are very different. Further extensions, of the nucleic acid modeling platform described herein are possible including the inclusion of protein homology modeling and *de novo* modeling. This approach allows ribonucleo-protein complexes to be accurately predicted.

V. Structure Based Sequence Alignment and Threading of Nucleic Acids.

The present invention provides a set of novel algorithms for aligning nucleic acid sequences. Traditional sequence alignment uses sequence similarity scoring matrices to perform alignment. Figure 7A shows the result of such an approach for 5S rRNA using CLUSTAL-W. The result shows that only 9% of the residues are correctly aligned. The reason for the failure is that RNA sequences conserve their secondary structure and the

identity of their single stranded loop regions. Thus, the base paired regions are not conserved at the sequence level but at a higher level (namely secondary structure). The algorithm of the present invention fully accounts for the nucleic acid secondary structure in the alignment process. This approach is called "structure based sequence alignment" (SBSA). Figure 7B shows the correct sequence alignment derived using the SBSA algorithm of the present invention. The structures shown in Figure 8 show that the SBSA algorithm of the present invention correctly placed nucleotides such that the correct secondary structure is represented. Importantly, sequence alignment by this method is equivalent to claiming that two nucleotides that are aligned occupy the same location in three-dimensional space. Thus this provides a method for determining the threading of a sequence into a known three-dimensional structure. This in turn is used for homology modeling (Section I, above).

```
A.  CLUSTAL-W sequence alignment of 5S rRNA
Identity Score = 68/122 = 55.7%        Correct alignment = 9/122 = 7.4%
1FFK    ----UUAGGCGGCCACAGCGGUGGGGUUGCCUCCCGUACCCAUCCCGAACACGGAAGAUA 56
1NKW    ACACCCCCGUGCCCAUAGCACUGUGGA-ACCACCCCACCCCAUGCCGAACUGGGUCGUGA 59
                    * * *** ***  ** **   ** ***   ***** ******  **  *  *

1FFK  AGCCCACCAGCGUUCCGGGGAGUACUGGAGUGCGCGAGCCUCUGGGAAACCCGGUU--CG 114
1NKW  AACACAGCAGCG--CCAAUGA-UACUCGGAC-CGCAGGGUCCCGGAAAAGUCGGUCAGCG 115
      * * ** *****  **    ** **** *     ***  *    * ** ***  ****   **

1FFK    CCGCCACC- 122
1NKW    CGGGGGUUU 124
          *  *

B.  SBSA Alignment of 5S rRNA
Score = 52/122 = 42.6%   Struc. Ident. = 109/122 = 89.3%    Correct alignment = 100%

                  ddd    RRRRRR d    RRRRRRRR   i   RRRRRRR               LLLL  LLL
1FFK              ---UUAGGCGGCC-ACAGCGGUGGGGUUGCCUCCCGUACCCAUCCCGAACACGGAAGAUAAG 59
1NKW              ACACCCCCGUGCCCAUAGCACUGUGGA-ACCACCCCACCCCAUGCCGAACUGGGUCGUGAAA 62
                    RRRRRRRRRR i    RRmRRRRR   d   RRRRRRR               LLLL  LLL

                  LLLLLL LL mmRRRRR i RRRRRRRRiRR    LL
1FFK              CCCACCAGCGUUCCGGGGAGUACUGGAGUGCGCGAGC
1NKW              CACAGCAGCGCCAAUGAU-ACUCGGAC---CGCAGC-
                  LLLLLm LL RRmRRRR d RRRRRRRdddR    Ld

                  LLLLLLLL  LLLLLmm ddLLLLLL   ddd
1FFK              CUCUGGGAAACCCGGUUC--GCCGCCACC--- 122
1NKW              -GUCCCGGAAAAGUCGGUCAGCGCGGGGGUUU 124
                  dLLLLLLL  LLLLmLL iiLLLLLLLL iii
```

Figure 7: Comparison of the structure alignments obtained by **(A)** CLUSTAL-W and **(B)** SBSA for the 5S rRNA of *H. marismortui* (1FFK) and *D. radiodurans* (1NKW). Note that CLUSTAL-W correctly aligns only 7.4% of the residues, while SBSA gets 100% correct. The code above and below sequence is as follows: i = insertion, d = deletion, R = nucleotide that is paired to its 3'-side, L = nucleotide that is paired to its 5'-

side, m = a nucleotide that participates in a mismatch that is known from phylogeny to be commonly replaced by base pairs. "Struc. Indent." means that the nucleotides are in identical locations in the secondary structures – note that deletions and insertions are scored as NOT structurally identical (there are 13 indels in the alignment above). In the CLUSTAL-W alignment, asterisks (*) indicate positions with identical residues, but this does not imply that they are correctly aligned. The results clearly indicate that for RNA, sequence identity alone is not sufficient to deduce proper alignment. Similar observations also apply to DNA and other strongly coded polymers.

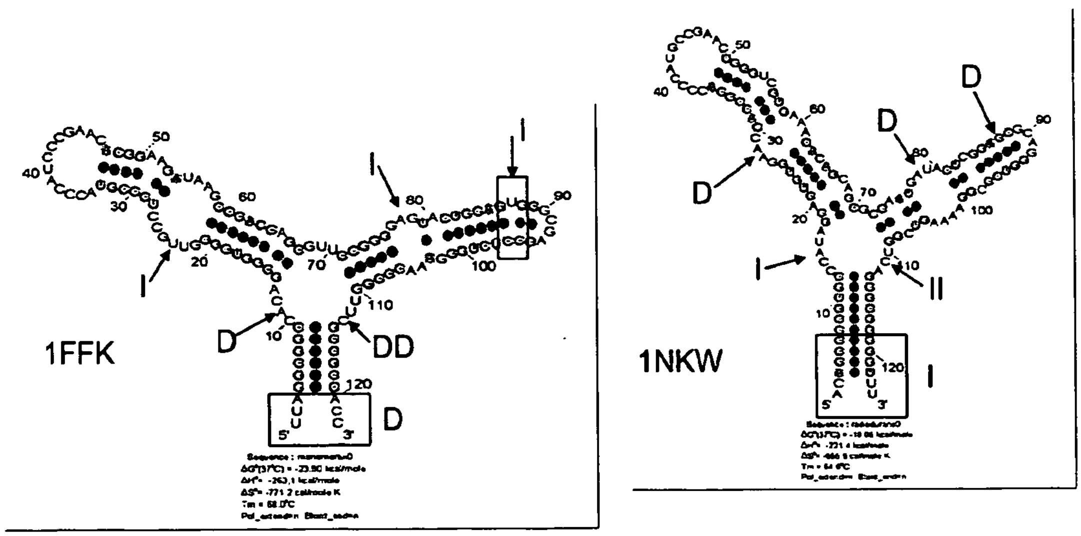


Figure 8: Secondary structures of the 5S rRNA sequences that are aligned in Figure 7. The positions of insertions and deletions are shown. The secondary structures were derived by comparative sequence analysis (Gutell ref) and the figures generated using Visual OMP. Insertion and deletion sites found in the SBSA from Figure 7 are indicted.

The SBSA can use a secondary structure from the reference template that is derived either from phylogeny, experimental methods, or from analysis of a three-dimensional structure (*e.g.*, as in Section II) and align the unknown sequence against the reference secondary structure. Alternatively, if the reference secondary structure is unknown the algorithm can perform an alignment of the two sequences subject to the constraint that they form the same secondary structure. This method may be less reliable than the one with where one secondary structure is known, but is useful in cases where only a few related sequences are known and may be a useful starting point for manual refinement of the alignment.

The SBSA starts by analyzing the reference template. There are three different preferred implementations of SBSA: 1. The secondary structures are known for both reference and query sequences, 2. The secondary structure is only known for the reference template, and 3. The secondary structure is unknown for both reference and query sequences. The alignments obtained are most reliable for case 1, and least reliable for case 3. Note that the final alignments may be manually optimized to insure proper placement of gaps, before inputting into the homology (Section I) or the de novo (Section II) algorithms.

Case 1: The secondary structures are known for both reference and query sequences (from phylogenetic analysis or from analysis of the three-dimensional structure). The algorithm starts by creating new strings for both sequences in which the nucleotide participating in pairs (both matches and mismatches) are replaced by the letters R and L. The resulting strings are called the "edited reference" and "edited query" strings. This novel process of editing the paired residues prevents incorrect alignment of the paired residues and yet retains their positions and allows for correct placement of insertions and deletions in the loop regions (*i.e.*, unpaired or single stranded regions). Next, the edited reference and edited query strings are aligned using a Needleman-Wunsch type global alignment algorithm (*ref) using the scoring matrix in Table 2, below. Note that the scoring matrix favorably scores nucleotide identities (A-A, C-C, G-G, and U-U) and pairs (R-R and L-L), but heavily penalizes R-L and L-R substitutions. In addition, transition mutations (A-G, G-A, C-T, T-C) are scored more favorably than transversion mutations (A-C, C-A, A-U, U-A, C-G, G-C, G-U, U-G), because conservation of purine or pyrimidine is often functionally important due to size and stacking considerations and thus evolutionarily conserved. The method described can use different scoring matrices, which may be optimized for specific classes of RNAs (tRNA, 16S rRNA, 23S rRNA, etc.). The alignment is achieved by a dynamic programming algorithm in which two alignment matrices, M and M' are created. M is the alignment matrix with terminal gap penalties = internal gap penalty and M' is the matrix in which terminal gaps penalty is set to zero (but internal gaps are still penalized by way of the gaps found in M(i, j), M(i, j-1), M(i-1, j). The elements M(i, 0) and M(0, j) are initialized to zero for all i and j. The elements in the alignment matrices, M(i, j) and M'(i, j), are

optimized according to the following recursive equations (this is called the "fill algorithm"):

```
M(i, j) = max{M(i-1, j-1) + S(Xi, Xj),                    //with terminal gaps penalized
          M(i-1, j) + W,
          M(i, j-1) + W}

M'(i, j) = max{ M(i-1, j-1) + S(Xi, Xj), M(i, j-1), M(i-1, j)}   //terminal gap has no penalty.
```

Where M(i, j) is the score for the fragment 1 to i of the reference sequence and fragment 1 to j of the query sequence. Xi is the identity of the nucleotide at position i in the reference string and Xj is the identity of the nucleotide at position j in the query string. S(Xi, Xj) is the substitution score from Table 1. W is the gap penalty which is given by the affine equation:

W = gap opening + (n-1)*gap extension
Where n is the number of nucleotides in the gap.

The sequence alignment is then obtained by the usual traceback algorithm using the M and M' matrices. Alternative embodiments of the alignment algorithm are to use a local alignment algorithm like Smith-Waterman or to use dot plot optimization using scoring matrices similar to Table 2.

Case 2: The secondary structure is only known for the reference template. The alignment proceeds initially identically to case 1, with the reference template edited to replace paired residues with R and L characters. The edited reference string is then aligned with the unedited query string using the fill algorithm from case 1 and the substitution matrix shown in Table 2. Next, the pattern of RRR and LLL stretch is analyzed to determine the hierarchy of pairing so that a tree diagram may be determined. The roots of the hierarchal tree have the largest difference between nucleotide positions and the leaves have the smallest difference and the branches have intermediate differences between nucleotide positions. The tree from the reference sequence is then superimposed on the query sequence using the preliminary alignment. Base pairs in the

query string are then confirmed (*e.g.*, does and R-L pair found in the reference correspond to a Watson-Crick or G-U or, UG pair in the query) in the order specified by the tree (*e.g.*, start with hairpins, then in decreasing order Watson-Crick pairs, mismatches, internal loops, bulges, multiloops, and bifurcations last). The query may have more or less pairs then found in the reference, and thus the algorithm proceeds by checking i+1, j-1 in the query until a mismatch is found. The algorithm only keeps the length of the paired region corresponding to the maximum length found in the reference (this prevents accidental over extension of the stems which would compromise the subsequent steps of the algorithm). A combinatorial number of paired alignments can be contemplated in which the paired regions are "slipped" with respect to one another or gaps are present. The first slipped alignment of pairs to consider is the one specified by the current sequence alignment. Alternative alignments are then considered in the order in which slipped orientations with the minimum number of gaps are considered first. If the query is found to contain a consecutive series of pairs, then a new edited query string is created in which the query has paired residues replaced by R and L. The new string is then realigned with the edited reference and the new alignment is used to generate a new tree. The process is repeated until all paired regions in the tree have been examined. At this point, proper alignment of most of the paired regions and the conserved unpaired regions are obtained. In the edited query string, the paired positions are used to generate a list of pairing constraints and the matched aligned loop regions are used to generate a list of unpaired constraints. These constraint lists are used in the thermodynamic DPA (*e.g.*, Visual OMP) to obtain the remaining pairs. The new pairs discovered by the thermodynamic DPA are then used to obtain a final edited query string with all R and L substitutions present. The final edited query string is then aligned against the edited reference string to obtain the final sequence alignment. It should be appreciated that there are alternative alignment strategies for aligning a sequence against a known secondary structure consensus.

Case 3. The secondary structure is unknown for both reference and query sequences. This proceeds using an iterative procedure similar to that described for case 2. This method, however, does not have a guide secondary structure for the reference string. Thus, the secondary structure of both sequences are obtained iteratively. The

algorithm starts by aligning the two sequences using the unedited reference and query strings using the algorithm from case 1 and substitution matrix in Table 2. The longest substring with the highest score is then constrained to be single stranded in the thermodynamic DPA. The two secondary structures obtained are then compared and all the pairs found in the two structures that are identical are kept and changed to R and L in edited reference and edited query strings. These are then aligned and the next longest and highest scoring substring is identified. This new string is then constrained to be single stranded along with the previously determined single stranded regions. The sequences are then repeatedly folded by the thermodynamic DPA and by the sequence alignment algorithm, until the thermodynamic algorithm generates secondary structures for both reference and query string that have the same tree diagram (order of LLL and RRR stretches). The final list of pairs is then used in to edit the reference and query strings with R and L to obtain a final sequence alignment. Note that during the iterative process, it is possible that residues that were once paired might become unpaired or vice versa; this prevents the overall case 3 algorithm from getting stuck in a local minimum. Also note that the case 3 algorithm is less reliable than case 2 or case 1, and thus manual refinement of the alignment may be desired. Alternative, structure based alignment algorithms have been proposed in the literature (DynAlign ref), but used substitution matrices that are not as effective as the one shown in Table 2. Note that occasionally phylogenetic multiple sequence alignments reveal mismatches that are replaced in other organisms as a Watson-Crick pair or other mismatch. Such mismatch cases are scored similarly to the R and L scoring given in Table 2.

Table 2: Substitution Scoring Matrix for Structure Based Sequence Alignm nt

		Top Sequence					
		A	C	G	U	R	L
Bottom Sequence	A	1	0.1	0.5	0.1	0.1	0.1
	C	0.1	1	0.1	0.5	0.1	0.1
	G	0.5	0.1	1	0.1	0.1	0.1
	U	0.1	0.5	0.1	1	0.1	0.1
	R	0.1	0.1	0.1	0.1	2	-2
	L	0.1	0.1	0.1	0.1	-2	2

Gap Opening =	-0.5
Gap Extension =	0
Terminal Gap =	0

VI. Identification of Conserved Hydrogen bonds.

The BP_GEOM algorithm determines all hydrogen bonds in a nucleic acid structure. These hydrogen bonds are then analyzed by BP_GEOM to determine the nucleotides participating in Watson-Crick base pairs, mismatches, base-backbone interactions, and backbone-backbone interactions. In confirming the presence of a Watson-Crick pair, the co-planarity of the participating bases is computed. The equation of the plane of a base i is determined from the XYZ coordinates of C6-N1-C2 of purines (A and G) and from C4-N3-C2 of pyrimidines (C, U, and T). The minimum distance between the plane of one base and the point N3 or N1 of the paired base is determined by standard geometry methods. If the distance is less than 2 Å and all H-bonds are less than 3.5 Å (distances for O6-N4, N1-N3, and N2-O2 for G-C pairs and N6-O4 and N1-N3 for A-U pairs), a putative base pair is declared to be a Watson-Crick pair. The co-planarity constraint avoids the output of false positive base pairs that are highly buckled or twisted, consecutive in the sequence, or stacked in the structure. The H-bond and co-planarity thresholds may be set to tighter or looser values if desired.

The algorithm also identifies if the H-bonds found in the reference template are also present (*i.e.*, conserved H-bond) in the homology model sequence (Section I). An H-bond is considered "conserved" if a D-A H-bond pair is found at the same XYZ coordinates within a defined error tolerance (*e.g.*, typically 3.0 Å) for both the reference structure and the homology model of the query sequence. These are used in the minimization (or molecular dynamics) of the homology model as constraints to maintain

the important interactions but to minimize the energy of less important nucleotides. The algorithm is also useful for extracting the secondary structure from a three-dimensional structure. The secondary structure is then useful for performing the SBSA (Section V). Note that the secondary structure derived by this method is more reliable than the phylogenetic secondary structure since it represents a single state of the nucleic acid rather than a superposition of states as is present in the phylogenetic structure (*i.e.*, the phylogenetic structure is the superposition of all structural states that are required for function). The algorithm then determines those Watson-Crick pairs that are part of the secondary structure and those that result from pseudoknots. The pseudoknots are identified by computing the number of chord crossings for all pairs in a Nussinov plot (by checking the conditions for i-j and k-l pairs: if $i<k<j<l$ or $k<i<l<j$ then a chord crossing is identified), and iteratively removing base pairs with the highest number of chord crossings from the base pair list.

A summary of certain preferred embodiments of the systems and methods are shown below, in Figure 9.

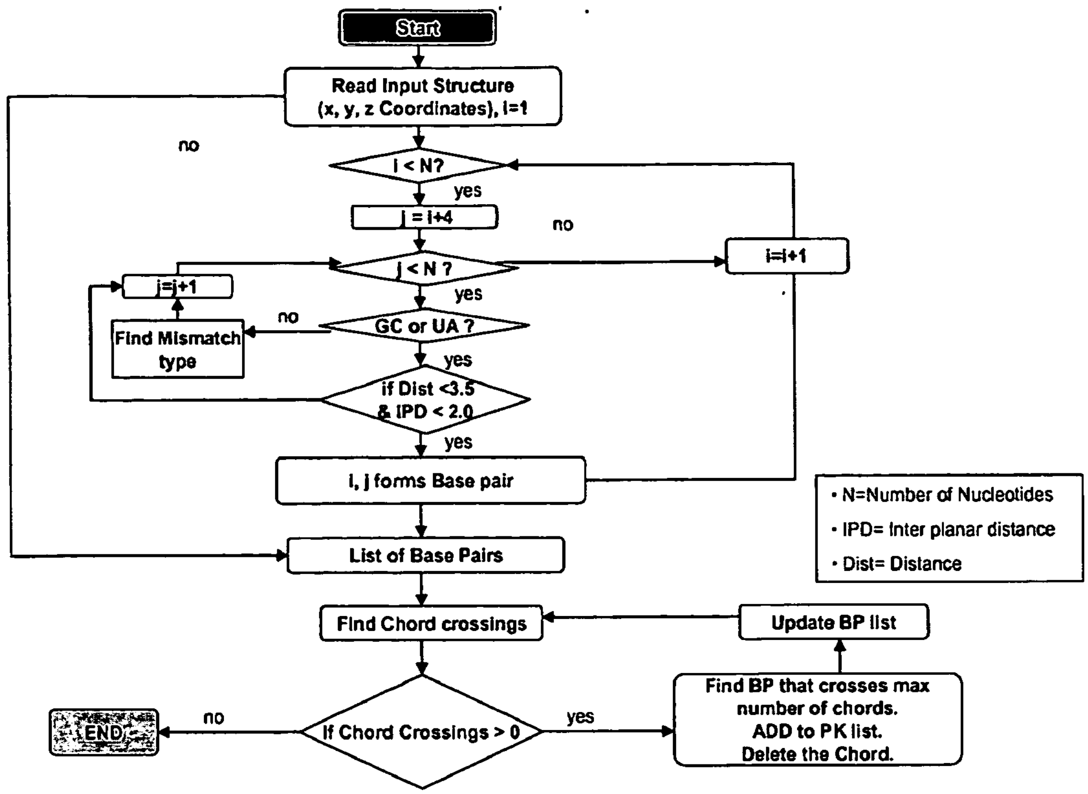


Figure 9: Flowchart of the BP_GEOM algorithm.

VII. Systems of the Present Invention

The methods of the present invention are implemented in a wide variety of systems and settings. In some preferred embodiments, the methods are conducted using software run on a computer processor to carry out the algorithms. While in preferred embodiments, the methods are carried out in an automated format, entirely within a computer processor, it should be understood that one or more components may be carried out be a human and that the methods may involve human interaction or intervention at one or more points.

The computer processor for conducting the methods of the present invention may be housed in any type of device, including, but not limited to, desktop computers, scientific instruments, hand-held devices, personal digital assistants, phones, medical instruments, implanted devices (*e.g.* in vivo), and the like. The methods need not be carried out on a single processor. For example, one or more steps may be conducted on a first processor, while other steps (simultaneously or sequentially) are conducted on a

second processor. The processors may be located in the same physical space or may be located distantly. In some such embodiments, multiple processors are linked over an electronic communications network (*e.g.*, an Internet).

In some preferred embodiments, the processors are associated with a display device for showing the results of the methods to a user or users. In some embodiments, the results comprise a video image of a predicted structure. In some embodiments, the results comprise coordinates of atoms, molecules, or motifs. In some embodiments, the results comprise a yes/no type answer to a specific question (*e.g.*, for medical diagnostic tests).

The processors of the present invention may also be directly or indirectly associated with information databases. In some embodiments, the databases comprise sequence information (*e.g.*, public or private nucleic acid sequence databases such as GENBANK). In some embodiments, the databases comprise structure databases, such as those described above. In yet other embodiments, the database comprises information pertaining to drugs, medical conditions, and/or patient-specific information.

EXPERIMENTAL

The following examples serve to illustrate certain preferred embodiments and aspects of the present invention and are not to be construed as limiting the scope thereof.

EXAMPLE 1

This example describes the use of the systems and methods of the present invention to provide improved structure analysis compared to the previously available methods. There are several examples in the literature of failed attempts to predict tRNA structure (Hubbard, J. M. & Hearst, J. E. (1991). Biochemistry 30, 5458-5465). Thus, prediction of tRNA tertiary structure is particularly suited to demonstrate the efficacy of the methods of the present invention.

To illustrate embodiments of the present invention, a number of methods were used: manual sequence alignment, methods of the present invention described in Section I for nucleotide substitution ("threading method"), and methods of the present invention described in Section II for *de novo* prediction ("*de novo* method") for a limited manually

constructed database of structural motifs based on only four structures (tRNAphe: 1EHZ, group I intron: 1HR2, *T. thermophilus* 30S ribosome: 1J5E, and *H. marismortui* 50S ribosome: 1JJ2). The results for the threading and *de novo* methods are shown in Figures 10 and 11. The all-atom RMSD for experimental vs. predicted structures are 2.4 and 3.8 Å, respectively for threading and *de novo* prediction. These predictions represent improvements over the previous history of failed attempts at tRNA structure prediction.

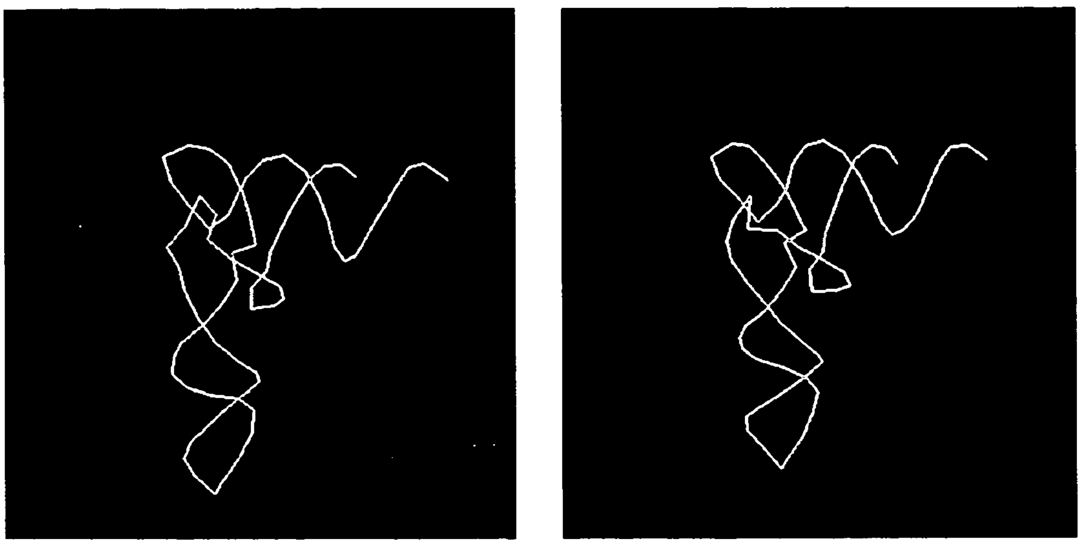


Figure 10: Left panel is the backbone of the x-ray structure of the human tRNAlys (PDB: 1FIR) with modifications removed. Right panel is the predicted structure of human tRNAlys obtained by the threading method with the human tRNAlys sequence into the yeast tRNAphe structure (PDB: 1EHZ) without AMBER refinement. The all-atom RMSD for the experimental vs. predicted structures is 2.4 Å (both structures have the chemical modifications removed). The sequences of human tRNAlys and yeast tRNAphe are 63% identical with no insertions or deletions.

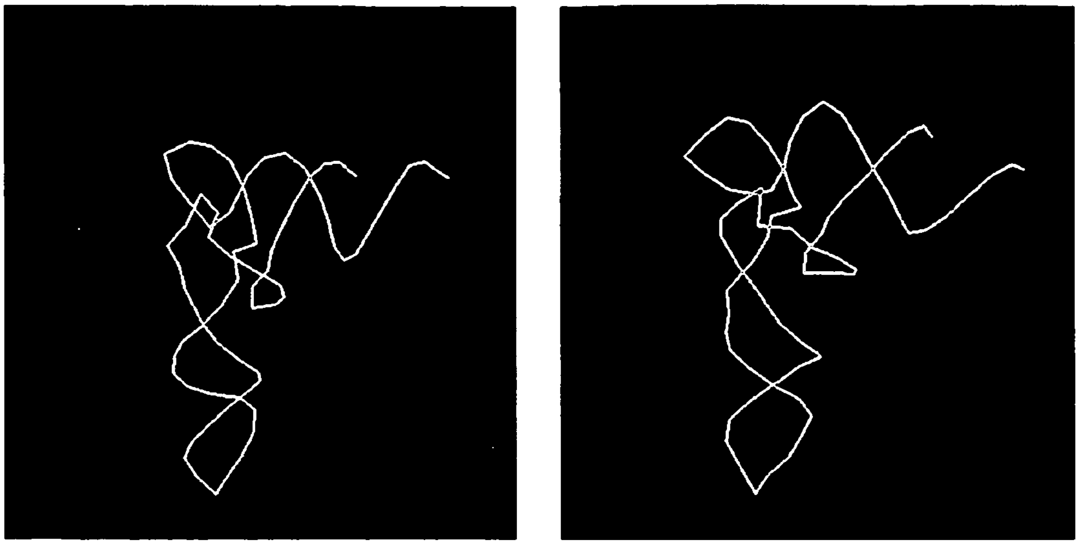


Figure 11: Left panel is the backbone of the x-ray structure of the human tRNAlys (PDB: 1FIR) with modifications removed. Right panel is the *de novo* method three-dimensional structure prediction made by the methods of Section II given the correct secondary structure, but no other information. The all atom RMSD compared to the experimental structure is 3.8 Å. For clarity only the backbone phosphorus atoms are shown.

EXAMPLE 2

This Example describes the use of the systems and methods of the present invention for design of therapeutic molecules. Knowledge of the structure of pathogen ribosomes is important for development of new narrow-spectrum (species-selective) antibiotics as well as broad-spectrum antibiotics. According to the CDC, the majority of hospital-acquired infections involve drug-resistant pathogens. Of particular concern are drug-resistant *Pseudomonas aeruginosa, Enterococcus, Escerichia coli, Staphylococcus aureus*, and *Mycobacterium turberculosis*. Development of new drugs against bioterrorism agents including *Bacillus anthracis, Francisella tularensis, Yersinia pestis*, and *Salmonella typhimurium* are particularly important in view of the risks of bioterrorism. Drug development would benefit highly from the availability of ribosome structures for different organisms.

Recently, a number of ribosome crystal structures that have been determined (9; 10; 11; 12; 13; 14). The ribosome is responsible for protein synthesis in all organisms.

About half of all clinically used antibiotics target the ribosome. Despite wide efforts, the only organisms that have had their ribosome structures determined at atomic resolution are the thermophilic bacterium *Thermus thermophilus* (high resolution of the 30S, and low resolution of the 70S), the acheaon *Haloarcula marismortui* (high resolution of 50S only), and the eubacterium *Deinococcus radiodurans* (high resolution of 50S only). Thus there is a need to predict the structures of other pathogenic prokaryotes and eukaryotes as well as the human ribosome. To date, software has not existed that could use the known ribosome structures to model the structures of ribosomes from homologous organisms and account for the substitutions, deletions, and insertions as well as chemical modifications in the ribosomes of other organisms.

The 16S ribosomal RNA contains several phylogenetically conserved subdomains that are believed to play fundamental roles in protein synthesis. One of the most highly conserved subdomains is commonly referred to as the "530 loop" (helix 18, nucleotides 500-540) – see Figure 12, below. The 530 loop has been implicated in tRNA binding (18; 19), translation fidelity (20; 21), and streptomycin resistance (22; 23; 24; 25). Recent crystal structures (9; 19) and numerous biochemical studies (18; 26; 27) have localized this loop to the site on the ribosome where information from the genome is decoded, known as the decoding region.

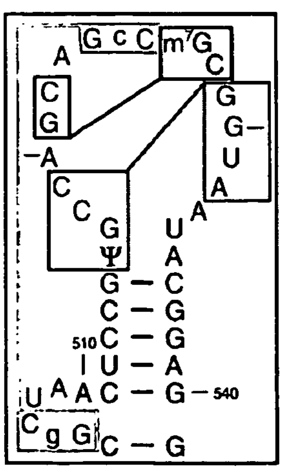


Figure 12: The 530 Loop of 16S rRNA. This loop contains a pseudoknot motif (yellow) as well as two modified nucleotides 7-methyl guanine at position 527 and pseudouridine at position 516, and is small enough to study by NMR spectroscopy. The use of model systems such as this allows rapid determinations of motif structures to generate a database per the systems and methods of Section III.

The 530 loop is ~~structurally unique~~. Phylogenetic analysis and the crystal structures have revealed interesting secondary and tertiary interactions within the 530 loop. Positions 521-522 are base paired with positions 527-528. This results in the formation of a tetraloop at positions 523-526. Positions 524-526 of the tetraloop interact with positions 505-507 in the stem bulge resulting in the formation of a pseudoknot 15; 17. In addition, two unique modified nucleotides are located in *E. coli* 16S RNA. One is a pseudouridine at position

516 and the other is a conserved 7-methylguanosine (m7G) at position 527. The pseudouridine is believed to serve primarily a structural role 28 but the role of the m7G is unclear.

Using the systems and methods of the present invention (*e.g.*, the systems and methods of Sections I and II), the structure of ribosomal subunits and portions of ribosomal subunits (e.g., the 530 loop) are determined from the many publicly available primary sequences. Motifs within these sequences are used to generate a database per the systems and methods of Section III. Having identified the structures, compounds are identified or selected that have a desired function in particular organisms. These compounds are then tested against organisms with similarly identified structures. In other embodiments, the identified structures are used to design rational therapeutics against classes of organisms that present similar structures. Additionally, in some embodiments, the presence of particular structures is used as a diagnostic characteristic to identify the presence of particular conditions or organisms and/or to select the appropriate intervention for such conditions or organisms.

Literature References:

1. Chow, C. S., Cunningham, P. R., Lee, K.-S., Meroueh, M., SantaLucia, J., Jr. & Varma, S. (2002). Photoinduced cleavage by rhodium complex at G-U mimsmatches and exposed guanines in large and small RNAs. Biochimie 84, 859-68.

2. Lee, K., Varma, S., SantaLucia, J., Jr. & Cunningham, P. R. (1997). In vivo determination of RNA structure-function relationships: analysis of the 790 loop in ribosomal RNA. J Mol Biol 269, 732-43.

3. Morosyuk, S. V., Cunningham, P. R. & SantaLucia, J., Jr. (2001). Structure and function of the conserved 690 hairpin in Escherichia coli 16 S ribosomal RNA: II. NMR solution structure. J. Mol. Biol., 197-211.

4. Morosyuk, S. V., Lee, K., SantaLucia, J., Jr. & Cunningham, P. R. (2000). Structure and function of the conserved 690 hairpin in Escherichia coli 16 S ribosomal RNA: analysis of the stem nucleotides. J Mol Biol 300, 113-26.

5. Morosyuk, S. V., SantaLucia, J., Jr. & Cunningham, P. R. (2001). Structure and function of the conserved 690 hairpin in Escherichia coli 16 S ribosomal RNA: III. Functional analysis. J. Mol. Biol., 213-228.

6. Meroueh, M., Grohar, P. J., Qiu, J., SantaLucia, J., Jr., Scaringe, S. A. & Chow, C. S. (2000). Unique Structural and Stabilizing Roles for the Individual Pseudouridine Residues in the 1920 Region of Escherichia coli 23S rRNA. Nucleic Acids Res 28, 2075-2083.

7. Consortium, I. H. G. S. (2001). Initial Sequencing and Analysis of the Human Genome. Nature 409, 860-921.

8. Venter, J. C. & al., e. (2001). The Sequence of the Human Genome. Science 291, 1304-1351.

9. Wimberly, B. T., Brodersen, D. E., Clemons, W. M., Jr., Morgan-Warren, R. J., Carter, A. P., Vonrhein, C., Hartsch, T. & Ramakrishnan, V. (2000). Structure of the 30S ribosomal subunit. Nature 407, 327-39.

10. Ban, N., Nissen, P., Hansen, J., Capel, M., Moore, P. B. & Steitz, T. A. (1999). Placement of protein and RNA structures into a 5 A-resolution map of the 50S ribosomal subunit. Nature, 400, 841-847.

11. Cate, J. H., Yusupov, M. M., Yusupova, G. Z., Earnest, T. N. & Noller, H. F. (1999). X-ray crystal structures of 70S ribosome functional complexes. Science, 285, 2095-2104.

12. Yusupov, M. M., Yusupova, G. Z., Baucom, A., Lieberman, K., Earnest, T. N., Cate, J. H. & Noller, H. F. (2001). Crystal Structure of the Ribosome at 5.5 A Resolution. Science 292, 883-896.

13. Ban, N., Nissen, P., Hansen, J., Moore, P. B. & Steitz, T. A. (2000). The Complete Atomic Structure of the Large Ribosomal Subunit at 2.4A Resolution. Science 289, 905-921.

14. Harms, J., Schluenzen, F., Zarivach, R., Bashan, A., Gat, S., Agmon, I., Bartels, H., Franceschi, F., Yonath, A. (2001). High Resolution Structure of the Large Ribosomal Subunit from a Mesophilic Eubacterium. Cell 107, 679-688.

15. Pleij, C. W., Rietveld, K. & Bosch, L. (1985). A new principle of RNA folding based on pseudoknotting. Nucleic Acids Res 13, 1717-31.

16. Pleij, C. W. (1995). Structure and function of RNA pseudoknots. Genet Eng (N Y) 17, 67-80.

17. Dam, E., Pleij, K. & Draper, D. (1992). Structural and functional aspects of RNA pseudoknots. Biochemistry 31, 11665-76.

18. Moazed, D. & Noller, H. F. (1990). Binding of tRNA to the ribosomal A and P sites protects two distinct sets of nucleotides in 16 S rRNA. J Mol Biol 211, 135-45.

19. Ogle, J. M., Brodersen, D. E., Clemons, W. M., Jr., Tarry, M. J., Carter, A. P. & Ramakrishnan, V. (2001). Recognition of cognate transfer RNA by the 30S ribosomal subunit. Science 292, 897-902.

20. Shen, Z. H. & Fox, T. D. (1989). Substitution of an invariant nucleotide at the base of the highly conserved '530-loop' of 15S rRNA causes suppression of yeast mitochondrial ochre mutations. Nucleic Acids Res 17, 4535-9.

21. O'Connor, M., Goringer, H. U. & Dahlberg, A. E. (1992). A ribosomal ambiguity mutation in the 530 loop of E. coli 16S rRNA. Nucleic Acids Res 20, 4221-7.

22. Santer, M., Santer, U., Nurse, K., Bakin, A., Cunningham, P., Zain, M., O'Connell, D. & Ofengand, J. (1993). Functional effects of a G to U base change at position 530 in a highly conserved loop of Escherichia coli 16S RNA. Biochemistry 32, 5539-47.

23. Powers, T. & Noller, H. F. (1991). A functional pseudoknot in 16S ribosomal RNA. Embo J 10, 2203-14.

24. Melancon, P., Lemieux, C. & Brakier-Gingras, L. (1988). A mutation in the 530 loop of Escherichia coli 16S ribosomal RNA causes resistance to streptomycin. Nucleic Acids Res 16, 9631-9.

25. Santer, U. V., Cekleniak, J., Kansil, S., Santer, M., O'Connor, M. & Dahlberg, A. E. (1995). A mutation at the universally conserved position 529 in Escherichia coli 16S rRNA creates a functional but highly error prone ribosome. Rna 1, 89-94.

26. O'Connor, M., Thomas, C. L., Zimmermann, R. A. & Dahlberg, A. E. (1997). Decoding fidelity at the ribosomal A and P sites: influence of mutations in three different regions of the decoding domain in 16S rRNA. Nucleic Acids Res 25, 1185-93.

27. Brimacombe, R. (1992). Structure-function correlations (and discrepancies) in the 16S ribosomal RNA from Escherichia coli. Biochimie 74, 319-26.

28. Lee, K., Holland-Staley, C. A. & Cunningham, P. R. (2001). Genetic approaches to studying protein synthesis: effects of mutations at Psi516 and A535 in Escherichia coli 16S rRNA. J Nutr 131, 2994S-3004S.

29. Rivas, E. & Eddy, S. R. (1999). A Dynamic Programming Algorithm for RNA Structure Prediction Including Pseudoknots. J. Mol. Biol. 285, 2053-2068.

30. Lee, K., Holland-Staley, C. A. & Cunningham, P. R. (1996). Genetic analysis of the Shine-Dalgarno interaction: selection of alternative functional mRNA-rRNA combinations. Rna 2, 1270-85.

31. M. Zuker & D. Sankoff. RNA Secondary Structures and their Prediction. *Bull. Mathematical Biology* 46, 591-621 (1984)

32. (2002). Protein Structure Prediction: A Bioinformatic Approach. Biotechnology Series (Tsigelny, I. F., Ed.), IUL, La Jolla.

34. Mathews, D. H., Sabina, J., Zuker, M. & Turner, D. H. (1999). Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J Mol Biol 288, 911-40.

35. Gautheret, D., Major, F. & Cedergren, R. (1993). Modeling the three-dimensional structure of RNA using discrete nucleotide conformational sets. J. Mol. Biol. 229, 1049-1064.

36. Zuker, M. (1989). On finding all suboptimal foldings of an RNA molecule. Science 244, 48-52.

All publications and patents mentioned in the above specification are herein incorporated by reference. Various modifications and variations of the described methods and systems of the invention will be apparent to those skilled in the art without departing from the scope and spirit of the invention. Although the invention has been described in connection with specific preferred embodiments, it should be understood that the invention as claimed should not be unduly limited to such specific embodiments. Indeed, various modifications of the described modes for carrying out the invention that are obvious to those skilled in the relevant fields are intended to be within the scope of the following claims.

CLAIMS

We Claim:

1. A system for generating a corrected three-dimensional model of nucleic acids, comprising a processor configured to:

   a) generate an initial, uncorrected model of a test sequence by comparison to a reference sequence;

   b) align secondary structure constraints of a reference sequence with a test sequence to generate an aligned sequence;

   c) make substitutions, deletions, and insertions dictated by said aligned sequence using geometrical computation algorithms for said substitutions and using molecular mechanics and molecular dynamics algorithms to close gaps caused by said deletions and insertions;

   d) identify conserved hydrogen bonds present in both said reference sequence and said uncorrected model to select hydrogen bond constraints; and

   e) optimize said uncorrected model using a forcefield algorithm that accounts for said hydrogen bond constraints to generate a corrected three-dimensional model of said test sequence.

2. A method for generating a corrected three-dimensional model of nucleic acids, comprising the step of submitting a test sequence to the system of claim 1 under conditions such that a corrected three-dimensional model of said test sequence is generated.

3. A system for predicting nucleic acid three-dimensional structure, comprising a processor configured to:

   a) compute a plurality of secondary structures of a test nucleic acid;

b) decompose said secondary structures into nucleic acid structure motifs;

c) rank said structure motifs in a hierarchal tree;

d) identify candidate three-dimensional motif structures for said motifs from a database of known three-dimensional structure motifs;

e) link said candidate three-dimensional motif structures in an order specified by said hierarchal tree to generate a candidate three-dimensional composite structure;

f) submit said candidate three-dimensional composite structure to an energy minimization algorithm to generate a refined candidate three-dimensional structure;

g) select a refined candidate three-dimensional structure based on best calculated energy to predict a three-dimensional structure of said test nucleic acid.

4. A method for generating a three-dimensional structure of a test nucleic acids, comprising the step of submitting a test sequence to the system of claim 3 under conditions such that a three-dimensional structure of said test sequence is generated.

5. A system for generating a nucleic acid structure motif database, comprising a processor configured to:

a) receive nucleic acid physical structure information;

b) decompose said physical structure information into nucleic acid structure ~~motifs~~;

c) associate data with said structure motifs, said data comprising: type of motif, size of motif, coordinates of backbone, and dihedral angles for bases;

d) compare said nucleic acid structure motifs to existing motifs in said database; and

e) add said structure motif and associate data to said database.

6. A method for generating a nucleic acid structure motif database, comprising the step of submitting nucleic acid physical structure information to the system of claim 5.

7. A system for refining nucleic acid structure predictions, comprising a processor configured to:

a) calculate energy minimization terms for a test nucleic acid structure prediction model, said energy minimization terms comprising: bond stretching, bond angles, torsion tress, and non-bonded internations;

b) optimize force constants, distance dependence, partial charges, and van der Waals radii parameters;

c) account for gap penalties for insertions or deletions, if present in said prediction model;

d) account for one or more experimental constraints associated with said test nucleic acid, said experimental constraints comprising hydrogen bonding information, nuclear Overhauser effect information, low resolution cryo-electron microscopy information, and chemical modification information;

e) employ distance constraints within a defined distance range but ignore distance constraints outside of said defined distance range; and

g) account for one or more nucleic acid folding thermodynamic measures, said nucleic acid folding thermodynamic measures comprising: folding entropy, solvation entropy.

8. A method for refining a nucleic acid structure prediction, comprising the step of submitting a nucleic acid structure prediction model to the system of claim 7.

ABSTRACT

The present invention relates to methods and systems for the accurate prediction of nucleic acid, *e.g.*, RNA and DNA, and other macromolecular and biomolecular three-dimensional structure from sequence and constraint information.

Document made available under the Patent Cooperation Treaty (PCT)

International application number: PCT/US04/037291

International filing date: 08 November 2004 (08.11.2004)

Document type: Certified copy of priority document

Document details: Country/Office: US
Number: 60/518,220
Filing date: 07 November 2003 (07.11.2003)

Date of receipt at the International Bureau: 15 December 2004 (15.12.2004)

Remark: Priority document submitted or transmitted to the International Bureau in compliance with Rule 17.1(a) or (b)